Proceedings
of the Conference
on

TELECOMMUNICATIONS IN CANADA:

Economic Analysis of the Industry

(MONTREAL, MARCH 4, 5, & 6, 1981)

Organized Jointly by
Ecole des Hautes Etudes Commerciales
and the
University of Victoria
in Collaboration with the
Department of Communications, Government of Canada

VOLUME I

1. Conference on Telecommunications in Canada: Economic Analysis of the industry (1981: Montreal)

Proceedings of the Conference on

2 TELECOMMUNICATIONS IN CANADA: ECONOMIC ANALYSIS OF THE INDUSTRY

(MONTREAL, MARCH 4, 5, and 6, 1981)

Organized Jointly By
Ecole des Hautes Etudes Commerciales
and the
University of Victoria
in Collaboration with the
Department of Communications, Government of Canada

VOLUME I

TELECOMMUNICATIONS IN CANADA
ECONOMIC ANALYSIS OF THE INDUSTRY

TABLE OF CONTENTS

IV. INTEGRATION OF FINANCIAL AND ECONOMIC ANALYSIS.

"The Value of the Firm Under Regulation and the Theory of the Firm Under Uncertainty: An Integrated Approach," Perrakis, Stylianas (University of Ottawa).

"Financing and Investment Behavior of the Regulated Firm Under Uncertainty," Berkowitz, Michael; and Cosgrove , E. (University of Toronto).

"Taxes, Financing and Investment for a Regulated Firm," Bernstein, Jeffrey (McGill University).

"Cost Based Tariffs, Integrated Network Use and Network Competition in the Telecommunication Sector," Müeller, Jürgen (CRG École Polytechnique).

Reviews and Discussions: Deschamps, Benoît (Georgia State University).

Author's Response: Berkowitz and Cosgrove (University of Toronto).

V. DEMAND ESTIMATION.

"Problems and Issues in Modelling Telecommunications Demand," Taylor, Lester (University of Arizona).

"Demande et consommation télephoniques: un modèle residential global," Curien, N., and Vilmin, E. (Direction Générale des Télécommunications, P.T.T. France).

"B.C.-Alberta Long Distance Calling," Lee, Marshall (British Columbia Telephone Company), and de Fontenay, Alain (Department of Communications, Government of Canada).

Comments: Nadiri, M. Isaq (New York University)
Sproule, Robert A. (Manitoba Telephone System)

Remarks: Dreesen, Erwin A. J. (British Columbia Telephone Company)

VI. LOCAL MEASURED SERVICES.

"Identifying Tariff-Induced Shifts in the Subscriber Distribution of Local Telephone Usage," Wong, T. Frank (Bell Laboratories).

"Local Telephone Costs and the Design of Rate Structures," Mitchell, Bridger (Rand Corporation).

"The Estimation of Usage Repression Under Local Measured Service--Empirical Evidence from the GTE Experiment," Wilkinson, G.F. (G.T.E.).

"Local Telephone Pricing: Two-Part Tariffs and Price Discrimination," Brander, James (Queen's University), and Spencer, Barbara (Boston College).

Comments: Hariton, George (C.R.T.C.)

Author's Response: Wong, T. Frank (Bell Laboratories)

VII. OPTIMAL PRICING.

"An Economic Analysis of Measured Service Options," Dansby, Robert (Bell Laboratories).

"The Regulatory Process Under Partial Information," Warskett, G. (Carleton University) and de Fontenay, Alain (Department of Communications, Government of Canada).

"Welfare Optimal Subsidy-Free Prices Under a Regulated Monopoly," Rhéaume, Gilles (Bell Canada).

"Efficiency, Equity and Regulation--A Model of Bell Canada," Breslaw, Jon, and Smith, J. Barry (Concordia University).

Discussion: Turton, D.O. (Canadian Pacific).

VIII. WELFARE IMPLICATIONS.

"The Welfare Effects of a Regulatory Constraint: A Productive Efficiency Approach," Diewert, W. (University of British Columbia).

"The Regulation of Telecommunication and Vertical Industry Structure," Westfield, Fred (Vanderbilt University).

"Empirical Evaluation of Cross-Subsidy Tests for Canadian Interregional Telecommunications Network," Autin, C, and LeBlanc, Gérald (Université Laval).

"Economic Evaluation of U.S. Telecommunications Policy Proposals," Rohlfs, J.H.; Goldstein, A.R.; and Marfisi, E.P. (American Telephone and Telegraph Company).

IX. NEW SERVICES.

"Rentabilité économique du trafic des nouveaux services sur les réseaux de télécommunications", de la Brunetiere, Jean (Direction Générale des Télécommunications, P.T.T. France).

"Strategic Analysis of Emerging Telecommunications Markets," Day, George; and Allison, William (University of Toronto).

X. COMPETITION, MONOPOLY AND VERTICAL INTEGRATION.

"An Overview of the Analysis of Competition vs. Monopoly in Telecommunication Services," Strick, John (Windsor University).

"Telesat Canada's Membership in Trans-Canada Telephone System: A Critique," Yale, Janet (University of Toronto).

INTRODUCTION

ELIZABETH KRIEGLER
DEPARTMENT OF COMMUNICATIONS

Ihe rapid technological change which characterizes the communications sector today is fuelled by the micro-electronics revolution and concurrent advances in satellite communications, digital transmission and switching and the potential development of fibre optics to name just a few. These technologies both complement each other in increasing the capacity and reach of telecommunications networks and compete with each other, not only in making possible a wider range of new services to both business and home than ever before in history, but in their contribution to the economic and industrial development of Canada and her ability to claim a share of the action in the international marketplace.

Innovation, productivity, price and quality are the key factors in that market place and therefore the primary concern of Canadian companies in the communications field. Creating an environment in which innovation can occur, productivity improvement is encouraged, and, price is reduced to a minimum while quality is enhanced is the concern and obligation of government. Creating that environment, is a particularly difficult task given the rapidly changing technologies, altering industrial structures, shifts in institutional arrangements and regulatory approaches during a decade when economic constraints are limiting the scope of governments to act. It can only be accomplished through the cooperation of all the players and a growing understanding by all of the many and changing variables that characterize the communications age in which we live

We have invited you to this conference to further this understanding, to exchange ideas and research, and the findings of that research; we are not here to indulge in a pure intellectual exercise, however stimulating such an exercise might be. We are here because our understanding of many fundamental economic conditions and interactions is sadly deficient. We are here because business men need that understanding to make crucial investment and marketing decisions and governmentsneed it to develop more realistic and flexible policies and regulatory approaches. We all have many questions and we are all groping in our attempts to make right and reasonable decisions. It is my hope that this conference will focus on the real and practical problems we face and that the combined wisdom of those presenting and discussing papers and those commenting from the floor will move us a little further towards an understanding of the many complex issues we face and therefore towards better and wiser decisions and policies.

E.C. Kriegler/mo

PRODUCTION ANALYSIS

A SURVEY OF RECENT RESULTS IN THE ANALYSIS OF PRODUCTION CONDITIONS IN TELECOMMUNICATIONS

MELVYN A. FUSS

University of Toronto

This paper surveys the recent empirical estimates of the telecommunications production technology which utilize flexible functional forms. Factor substitution, scale and technical change characteristics are analysed. Recent attempts to explore the issue of economies of scope in multi-output production processes are also analysed. One major finding of the survey is that decomposition of intertemporal efficiency gains between scale effects and effects due to technological change is very sensitive to functional form specification, even within the family of second order approximations. A second important finding is that scale elasticity estimates substantially above unity imply rates of technical change which seem to be unreasonably low when compared with estimates drawn from Canadian manufacturing.

1.0 Introduction

Recent advances in the econometric literature (utilizing the duality between cost and production) have made it possible to represent the telecommunications production process by a structure of technology sufficiently general so as to capture its important features. The purpose of this paper is to survey the empirical results emanating from the application of these relatively new techniques to telecommunications.

In general there are three characteristics of production which are of interest: (1) factor substitution possibilities, (2) output expansion (scale) effects, and (3) the rate and bias of technological progress. The second characteristic has traditionally been given the most attention in telecommunications since it is closely connected to the natural monopoly question. However estimation of the rate of technical change can also be an input into the natural monopoly issue in a dynamic context. Finally, biases in technical change and factor substitution characteristics have implications for capital accumulation and employment opportunities.

The studies analysed in this survey utilize second order flexible functional forms. The major advantage of these forms is their ability to represent multi-input, multi-output production processes characterized by variable elasticities of substitution and transformation, non-homothetic output expansion effects, and biased technical change. Earlier attempts to estimate the telecommunications production process employed functions which implied unitary, or at least constant elasticities of substitution and transformation, homogeneous expansion effects, and Hicks-neutral technical change (Dobell et al (1972), Mantell (1974), Vinod (1976), Waverman (1976)). The empirical results contained in papers reviewed in this survey demonstrate that the above restrictions are inappropriate for telecommunications. Formal tests carried out by Fuss and Waverman (1977), Denny et al (1979) and Nadiri and Shankerman (1979) confirm these less rigorous impressions.

2. Choice of Behavioural Model and Functional Form

2.1 Choice of the Behavioural Model

The empirical literature to be surveyed in this paper places the cost function at the centre of the estimation procedure.[1] The choice of the cost function rather than the production function as the basic building block is due to a number of advantages possessed by the cost function. Most important public policy issues in telecommunications require a knowledge of the cost structure, and the cost function is the most direct way of obtaining this needed information.[2] Since the observation unit is usually the individual firm, factor prices are likely to be more exogenous than factor inputs, reducing the possibility of simultaneous equations bias.[3] It is easier to specify a functional form which represents a sufficiently general technology using a cost function, particularly when output disaggregation is necessary. Finally, application of Shephard's Lemma provides a direct, simple way of generating a system of factor demand functions. Estimation of the demand system along with the cost function increases the number of observations without increasing the number of parameters. Generation of a system of factor demand functions from the production function is difficult unless constant returns to scale is imposed a priori.

The main disadvantage associated with estimation of the cost function is the need to assume cost minimizing behaviour. Since most telecommunications firms are monopolists in at least one of their service categories, competitive pressures cannot be relied on to force cost-minimizing input choices. In addition, investor-owned telecommunication firms are subject to rate of return regulation which may induce Averch-Johnson effects.[4]

The production of telecommunications services is a capital-intensive process, characterized by the use of physically long-lived capital-equipment much of which is "putty-clay" in nature. Hence a dynamic intertemporal cost-minimizing model with increasing marginal costs of adjustment is appropriate. Such a model has not been estimated in telecommunications. The question arises as to which of two polar cases provides the best approximation to this model: (1) a long-run constant marginal costs of adjustment model (the unrestricted cost function), or (2) a model which does not attempt to explain the time path of capital services and treats capital as exogenous (the restricted cost function). All the studies surveyed in this paper estimate the unrestricted cost function model. Denny et al (1979) and Christensen et al (1980) also consider the restricted cost function model. The former set of authors present a detailed conceptual comparison of the two models and conclude that for Canadian telecommunications the unrestricted cost function is the more appropriate approximation.

For investor-owned telecommunications firms, it is reasonable to assume present value (profit) maximizing behaviour, subject to regulatory constraint. Fuss and Waverman (1977) developed an econometric model in which the telecommunications firm chooses the profit maximizing levels of toll services but is constrained by the regulatory authorities to charge a price for basic local service below the profit-maximizing price. This restricted profit maximizing assumption results in an additional estimating equation for each toll service considered. The Fuss-Waverman model has been used by Denny et al (1979), Breslaw and Smith (1980), and Fuss and Waverman (1980). The last-named authors showed that the model could be obtained from a dynamic specification in which the objective is to maximize the intertemporal utility function of the investor-owners of the firm.

2.2 Choice of the Functional Form

All but one of the studies surveyed employ the translog second order approximation to an arbitrary cost function. The translog approximation to the cost function $C(Z_1 \ldots Z_n)$ takes the form

$$\log C(Z_1 \ldots Z_n) = \alpha_0 + \sum_{i=1}^{n} \alpha_i \log Z_i + \tfrac{1}{2} \sum_{i=1}^{n} \alpha_{ii} (\log Z_i)^2$$

$$+ \sum_{i \neq j} \sum \alpha_{ij} \log Z_i \log Z_j \qquad (1)$$

where the Z_i are the exogenous variables, usually factor prices, outputs and technology shift variables. In the case of the restricted cost function, one of the exogenous variables is capital. Nadiri and Shankerman (1979), Breslaw and Smith (1980), Denny and Fuss (1980), and Christensen et al (1980) all use the translog model in the form given by (1). Fuss and Waverman (1977) restrict technical change to be capital-augmenting, while Denny et al (1979) assume technical change is output-augmenting.

Fuss and Waverman (1980) estimated a generalization of the translog model which applies a Box-Cox transformation to the output levels. In this hybrid translog form, the output components of $\log Z_1 \ldots \log Z_n$ are replaced by

$$Q_j^* = \frac{Q_j^\theta - 1}{\theta} \qquad (2)$$

where Q_j is output j and θ is a parameter to be estimated. Since $\lim_{\theta \to 0} \frac{Q_j^\theta - 1}{\theta} = \log Q_j$, the hybrid translog cost function can be used to investigate the effects on the estimated cost structure of departures from the translog maintained hypothesis.

3. Estimation Procedures

The translog cost model is usually estimated as a systems multivariate regression model consisting of the cost function and (n-1) factor share equations, using the Zellner iterative estimation procedure. This approach was adopted by Nadiri and Shankerman (1979) and Christensen et al (1980). Denny and Fuss (1980) employed the two-stage estimation procedure suggested by Fuss (1977) for the case of a large number of inputs.

The inclusion of profit-maximizing behaviour adds "revenue share" equations to the system and renders outputs endogenous. In addition, estimates of service demand elasticities are necessary. Simultaneous equations estimation techniques were used by Fuss and Waverman (1977, 1980), Denny et al (1979), and Breslaw and Smith (1980) to overcome potential simultaneous equations bias. In addition, Breslaw and Smith (1980) and Fuss and Waverman (1980) estimated the factor demand and service demand systems together, incorporating the across-equations constraints implied by the presence of service demand elasticities in both systems. The result is a fully efficient estimation procedure for the Fuss-Waverman model.

Table 1 presents a summary comparison of the basic features, discussed above, of the studies of telecommunications production technology surveyed in this paper. While these studies differ in a number of details they all take as their starting point the duality theory between cost and production as embodied in the cost function.

Table 1

A Comparison of Basic Features of Studies of Telecommunication Production Technology

Features	Fuss-Waverman (1977)	Denny et al (1979)	Nadiri-Shankerman (1979)	Breslaw-Smith (1980)
Data Set	Bell Canada, 1952-75	Bell Canada, 1952-76	A.T.&T., 1947-76	Bell Canada, 1952-78
Outputs	Q_1 = message toll Q_2 = private line + WATS + TWX + miscellaneous Q_3 = local	Q_1 = message toll Q_2 = private line + WATS + TWX Q_3 = local + miscellaneous	single aggregate output	Q_1 = local Q_2 = message toll + WATS
Inputs	Aggregate capital, labour, materials	Aggregate capital, labour, materials	Aggregate capital, labour, materials, research and development	Aggregate capital, labour, materials
Functional Form	Translog Cost Function	Translog Cost Function	Translog Cost Function	Translog Cost Function
Technical Change Indicators	Capital augmenting time trend	Output augmenting access to direct distance dialing facilities and conversion to modern switching facilities - 2 indicators	Time trend	Index of switching and accessibility to the system
Behavioural Assumptions	Constrained profit maximization	Constrained profit maximization	Cost minimization	Constrained profit maximization
Method of Estimation	Iterative 3SLS - cost and demand systems estimated separately	Iterative 3SLS - cost and demand systems estimated separately	Iterative Zellner	Full Information Maximum Likelihood - cost and demand systems estimated simultaneously

Table 1 continued

Features	Denny-Fuss (1980)	Christensen et al (1980)	Fuss-Waverman (1980)
Time Period	Bell Canada, 1952-72	A.T.&T. 1947-77	Bell Canada, 1952-78
Outputs	Single aggregate output	Single aggregate output	Q_1 = message toll + WATS Q_2 = private line + TWX Q_3 = local + miscellaneous
Inputs	Aggregate capital, materials, 4 occupational labour groups	Aggregate capital, labour, materials	Aggregate capital, labour, materials
Functional Form	Two-stage translog cost function	Translog cost function	Hybrid Translog cost function
Technical Change Indicators	Access to direct distance dialing facilities	Distributed lag function of R+D expenditures by A.T.& T.	Output augmenting access to direct distance dialing facilities and conversion to modern switching facilities (2 indicators)
Behavioural Assumptions	Two-stage cost minimization	Cost minimization	Constrained profit maximization
Method of Estimation	Two-stage iterative Zellner	Iterative Zellner	Iterative 3SLS - cost and demand systems estimated simultaneously

4. Evidence on Factor Substitution

4.1 Measurement of Factor Substitution

The two most common measures of factor substitution effects are the constant output Allan-Uzawa (A-U) partial elasticity of substitution and the constant output cross-partial elasticity of demand. The A-U elasticity can be calculated from the cost function C as

$$\sigma_{ij} = \frac{CC_{ij}}{C_i C_j} \tag{3}$$

where C_i, C_j and C_{ij} are partial derivatives of the cost function. The cross-partial elasticity of demand for factor i with respect to a change in the price of factor j can be calculated as

$$\varepsilon_{ij} = \frac{\partial \log X_i}{\partial \log p_j} = S_j \cdot \sigma_{ij} \tag{4}$$

where S_j is the cost share of factor j.

I will use the constant output cross-partial elasticities to compare estimates of factor substitution. If $\varepsilon_{ij} > 0$ factors are substitutes, if $\varepsilon_{ij} < 0$, they are complements and if $\varepsilon_{ij} = 0$ they are independent. For the translog and hybrid translog cost functions, ε_{ij} can be calculated as

$$\varepsilon_{ij} = \frac{1}{S_i}[\gamma_{ij} + S_i S_j] \tag{5}$$

where $\gamma_{ij} = \dfrac{\partial^2 \log C}{\partial \log p_i \, \partial \log p_j}$

In addition, the own price elasticity of demand (output constant) can be calculated as

$$\varepsilon_{ii} = \frac{1}{S_i}[\gamma_{ii} - S_i + S_i^2] \qquad (6)$$

where $\gamma_{ii} = \dfrac{\partial^2 \log C}{\partial \log p_i^2}$

Table 2 contains a summary of the substitution characteristics of aggregate inputs as estimated by the various studies under review. Table 3 presents the own price elasticities of demand, estimated either at the mean of the data set or for a year at the midpoint of the data. Capital and labour and labour and materials are estimated as substitutes in production. Conflicting results have been obtained for capital and materials. However studies reporting both complementary and substitutability characteristics agree that the capital-materials input mix is the least responsive to relative price changes. In all cases the cross-price effects are statistically significant indicating that the choice of inputs by the telecommunications firm is responsive to changes in factor prices. Estimates of own price elasticities vary considerably across studies. Nevertheless there is general agreement that demands for aggregate inputs are inelastic and that the capital input is the least responsive to price changes. This latter fact is hardly surprising since capital equipment in this industry have long physical lifetimes and second hand markets are essentially non-existent.

The only evidence on factor substitution among disaggregated inputs is contained in a study by Denny and Fuss (1980). They disaggregated the labour category into operators, plant craftsmen, clerical workers and white

Table 2

Factor Substitution Characteristics - Aggregate Inputs

Factors	Fuss-Waverman (1977)	Denny et al (1979)	Nadiri-Shankerman (1979)	Breslaw-Smith (1980)	Fuss-Waverman (1980)
Capital-Labour	substitutes	substitutes	substitutes	substitutes	substitutes
Capital-Materials	substitutes	complements	substitutes	substitutes	complements
Labour-Materials	substitutes	substitutes	substitutes	substitutes	substitutes

Table 3

Own Price Elasticities of Demand - Aggregate Inputs*

Factors	Fuss-Waverman (1977)	Denny et al (1979)	Nadiri-Shankerman (1979)	Breslaw-Smith (1980)	Fuss-Waverman (1980)
Capital	-.671	.019 (.026)	-.26 (.04)	-.369	0**
Labour	-.989	-.397 (.044)	-.55 (.06)	-.773	-.437 (.033)
Materials	-1.02	-.541 (.087)	-1.12 (.13)	-.577	-.371 (.060)

* Standard errors in parentheses

** Constrained estimate. Unconstrained estimation yielded a positive price elasticity which was statistically insignificantly different from zero.

collar (executive and supervisory) personnel. All factors were found to be substitutes except operators and capital and clerical workers and plant craftsmen. Demands were inelastic except for operators and white collar personnel. The operators/capital complementarity is particularly interesting because a large number of studies have found unskilled labour and capital to be substitutes. However this result can be explained by the fact that technical change has led to a substitution of capital for operators which dominates the price complementarity (see section 6). This substitution relationship would have been attributed to a factor price effect if Hicks neutral technical change had been imposed, as was the case with previous studies.

5. Evidence on Scale Effects

Perhaps the most important production characteristic for policy purposes is the behaviour of costs as outputs vary; since this behaviour can establish whether or not a telecommunications firm such as Bell Canada or A.T.& T. is a natural monopoly over some range of its service offerings. Baumol (1977) has refined the definition of a natural monopolist and shown that the basic requirement is that the cost function be "sub-additive". A firm's cost function is sub-additive if it can produce any configuration of outputs at a lower cost than that attained by multi-firm production. Baumol shows that a firm may exhibit diseconomies of large-scale production and still be a natural monopolist under the sub-additivity definition, or, conversely if it produces more than one product may exhibit increasing returns to scale and still not be a natural monopolist. Hence the preoccupation with economies of scale to the exclusion of other output characteristics of the cost structure for the multi-product firm is misplaced.

The additional concept that needs to be considered is that of "economies of scope". A production technology exhibits economies of scope when for any configuration of multiple outputs, these outputs can be produced at less cost by a firm which operates a multi-product technology than if the same outputs were produced by a number of firms each operating a single product technology. While the necessary conditions for sub-additivity have yet to be established, Baumol demonstrates that the simultaneous existence of economies of scale and economies of scope are sufficient to insure sub-additivity.[6] Panzar and Willig (1977) have shown that a natural monopolist (defined in terms of a sub-additive cost function) may not be sustainable in the face of competitive entry in one of the multi-product monopolist's markets. A monopolist's pricing strategy is said to be sustainable if it can find a set of stationary product and quantity prices which does not attract rivals into the industry. Baumol, Bailey and Willig (1977) demonstrate that a natural monopolist (again defined as a multi-product firm with a sub-additive cost function) is sustainable if it chooses the Ramsey-optimal rate structure. The Ramsey-optimal rate structure is equivalent to the inverse elasticity rule for quasi-optimal pricing when the cross-price elasticities of demand for the multi-product firm's outputs are zero.

While sub-additivity is the basic cost concept of interest, it is very difficult to test per se. However the sufficient conditions for sub-additivity - economies of scale and economies of scope - are more amenable to the formulation of testable hypotheses. In this section we will survey the recent empirical results concerning economies of scale and scope. We begin with some necessary definitions of scale and scope in terms of characteristics of the cost function.

5.1 Tests of Overall Economies of Scale and Overall Economies of Scope

The starting point for testing the natural monopoly hypothesis is the construction of a test for overall (aggregate) economies of scale. Overall economies of scale exist if an increase in all outputs of $\lambda\%$ leads to a cost increase of less than $\lambda\%$. As shown by Panzar and Willig (1979) and Fuss and Waverman (1977), local overall economies of scale are measured by the scale elasticity

$$S = \frac{1}{\sum_{j=1}^{N} \varepsilon_{CQ_j}} \qquad (7)$$

where ε_{CQ_j} is the cost-output elasticity of the j-th output. If $S > 1$, economies of scale prevail locally, if $S < 1$ diseconomies of scale prevail and if $S = 1$, constant returns to scale prevail. Of course in the aggregate output specification, $N = 1$.

A global test of economies of scope can be formulated in the following way. Suppose an N output production process can be represented by the joint cost function

$$C = C(Q_1, Q_2, \dots Q_N) \tag{8}$$

where factor prices and any other arguments of the cost function have been suppressed for simplicity. Overall economies of scope can be determined by comparing the cost of producing each output separately (the "stand alone" cost) with the actual joint cost. The relevant expression is

$$SC = \sum_{j=1}^{N} C_j(Q_j) - C(Q_1, Q_2, \dots Q_N) \tag{9}$$

If $SC > 0$, economies of scope exist; if $SC < 0$, diseconomies of scope exist and independent production is cost-minimizing. If $SC = 0$, joint production neither yields cost savings nor causes cost increases.

It should be noted that to compute a global necessary and sufficient test of overall economies of scope requires that one be able to compute stand-alone costs. In telecommunications this would require observations on independent production of outputs such as message toll, competitive, and local services. Clearly, we do not have the required set of observations and hence a global test of overall economies of scope is not possible.

A local sufficient test of overall economies of scale is possible. Panzar and Willig (1977) have shown that

$$\frac{\partial^2 C}{\partial Q_i \, \partial Q_j} < 0 \qquad i,j = 1,\ldots N; \quad i \neq j \qquad (10)$$

is sufficient for the existence of overall economies of scope. However, as noted by Fuss and Waverman (1977) and Baumol, Fischer and Nadiri (1978), the local nature of this test makes it a very weak one. We conclude that there exists no satisfactory test of overall economies of scope due to data limitations.

5.2 Product-Specific Economies of Scope and Economies of Scale

One particular public policy issue of considerable importance is the question of whether competition in the provision of certain services should be encouraged. Researchers can shed light on this issue by attempting to estimate the extent of product specific economies of scope and economies of scale in the provision of private line services. One requirement for computing product-specific economies of scope is that one observe a production process in which a zero amount of the product under consideration is produced. For private line services in Canada this requirement is approximately met, since Bell Canada produced a very small output of this service in the early 1950's, which is part of the data sample. Unfortunately, a second requirement for computing private line-specific economies of scope is that one observe independent production of this output, so that stand-alone costs can be estimated. The hybrid translog cost function, unlike the ordinary translog function permits the estimation of stand-alone costs. However this estimation requires extrapolation of the cost function well outside the observed data points for toll and local services, and thus considerable caution must be exercised in interpreting the results.

The test for product-specific economies of scope is as follows. Suppose private line service is the j-th service output. Product-specific economies of scope with respect to private line service exist if

$$C(Q_1, Q_2, \ldots Q_{j-1}, 0, Q_{j+1}, \ldots Q_N) + C(0, \ldots 0, Q_j, 0, \ldots 0) - C(Q_1, Q_2, \ldots Q_N) > 0 \qquad (11)$$

Panzar and Willig (1979) have defined the degree of product specific economies of scope as

$$SC_j = \frac{C(Q_1,Q_2,\ldots Q_{j-1},0,Q_{j+1},\ldots Q_N) + C(0,\ldots 0,Q_j,0,\ldots 0) - C(Q_1,\ldots Q_N)}{C(Q_1,\ldots Q_N)} \qquad (12)$$

If $SC_j > 0$, SC_j measures the proportionate increase in cost from separating private line services from the production of other services. If $SC_j < 0$, it measures the proportionate cost decrease from independent production of private line services.

Panzar and Willig (1979) have also proposed a measure of product specific economies of scale. They define the degree of product j specific economies of scale as

$$S_j = \frac{IC_j}{Q_j \frac{\partial C}{\partial Q_j}} \qquad (13)$$

where $IC_j = C(Q_1,Q_2,\ldots Q_N) - C(Q_1,\ldots Q_{j-1},0,Q_{j+1},\ldots Q_N)$ is the incremental cost of producing product j . It can be shown that (13) can be written in the form

$$S_j = \frac{IC_j}{C} \Big/ \varepsilon_{CQ_j} \qquad (14)$$

If $S_j > 1$, there exists product j specific economies of scale (locally). If $S_j < 1$, there exists diseconomies of scale and if $S_j = 1$, there exists constant returns to scale.

One final test of product-specific returns to scale is of interest. Suppose private line services are produced by two firms in the amounts Q_2^1 and Q_2^2, so that industry output is $Q_2 = Q_2^1 + Q_2^2$. We are interested in whether the takeover of firm 2's output by firm 1 would allow firm 1 to produce the additional output under increasing returns to scale (declining average incremental cost). Fuss and Waverman (1980) have shown that the degree of returns to scale associated with this takeover can be computed as

$$\tilde{S}_2 = \left[\frac{IC_2}{C} \Big/ \varepsilon_{CQ_2}\right] \cdot \left[1 - \frac{Q_2^1}{Q_2}\right] \qquad (15)$$

where IC_2, C, and ε_{CQ_2} are all evaluated at Q_2. If $\tilde{S}_2 > 1$, then the additional production is subject to increasing returns to scale. If $\tilde{S}_2 < 1$ decreasing returns to scale prevail and if $\tilde{S}_2 = 1$ the additional production is subject to constant returns to scale. This final test has an obvious application to the Bell Canada-CNCP Interconnection case since it can be used to test whether the efficient market structure (in a static sense) is for Bell to become the monopoly supplier of private line services.

5.3 Testing for the Presence of Economies of Scale and Scope

For flexible functional forms measures of economies of scale and scope are functions of the data as well as the estimated parameters. Hence testing of hypotheses requires one to use the approximation methodology suggested by Denny and Fuss (1977). The usual procedure is to test a hypotheses at the mean by transforming (scaling) the data so that all variables equal unity at the mean observation. For the translog and hybrid translog models, the resulting test statistic is usually a function of parameter estimates alone, since $\log Z_i$ and $\frac{Q_j^\theta - 1}{\theta}$ are both zero when Z_i , $Q_j = 1$. We will illustrate the procedure for the translog and hybrid translog specifications used by Denny et al (1979) and Fuss and Waverman (1980). The translog specification used by Denny et al (1979) is

$$\begin{aligned}\log C = {} & \alpha_0 + \sum_i \alpha_i \log p_i + \sum_k \beta_k \log Q_k^* \\ & + \tfrac{1}{2} \sum_i \gamma_{ii} (\log p_i)^2 + \sum_{\substack{i \\ i \neq j}} \sum_j \gamma_{ij} \log p_i \log p_j \\ & + \tfrac{1}{2} \sum_k \delta_{kk} (\log Q_k^*)^2 + \sum_{k \neq \ell} \sum \delta_{k\ell} \log Q_k^* \log Q_\ell^* \\ & + \sum_i \sum_k \rho_{ik} \log p_i \log Q_k^* \end{aligned} \qquad (16)$$

where p_i is an input price, Q_k^* is a technical change augmented output (see section 6 for details), i,j are indexed over inputs and k, ℓ are indexed over outputs. The aggregate scale elasticity is given by

$$S = \frac{1}{\sum_\ell \varepsilon_{CQ_\ell}} = \left[\sum_\ell (\beta_\ell + \sum_k \delta_{k\ell} \log Q_k^* + \sum_i \rho_{i\ell} \log p_i) \right]^{-1} \qquad (17)$$

which reduces to $S = \left[\sum_\ell \beta_\ell \right]^{-1}$ at the transformed means.

The hybrid translog cost function used by Fuss and Waverman (1980)[7] is specified as

$$\log C = \alpha_0 + \sum_i \alpha_i \log p_i + \sum_k \beta_k \left[\frac{Q_k^{*\theta} - 1}{\theta}\right]$$

$$+ \tfrac{1}{2} \sum_i \gamma_{ii} (\log p_i)^2 + \sum_i \sum_{\substack{j \\ i \neq j}} \gamma_{ij} \log p_i \log p_j$$

$$+ \tfrac{1}{2} \sum_k \delta_{kk} \left[\frac{Q_k^{*\theta} - 1}{\theta}\right]^2 + \sum_k \sum_{\substack{\ell \\ k \neq \ell}} \delta_{k\ell} \left[\frac{Q_k^{*\theta} - 1}{\theta}\right] \left[\frac{Q_\ell^{*\theta} - 1}{\theta}\right]$$

$$+ \sum_i \sum_k \rho_{ik} \log p_i \left[\frac{Q_k^{*\theta} - 1}{\theta}\right] \qquad (18)$$

The aggregate scale elasticity is given by

$$S = \left[\sum_\ell \varepsilon_{CQ_\ell}\right]^{-1} = \sum_\ell \left[Q^{*\theta}_\ell \cdot \left\{ \beta_\ell + \sum_k \delta_{\ell k} \left[\frac{Q_k^{*\theta} - 1}{\theta}\right] + \sum_i \rho_{i\ell} \log p_i \right\} \right]^{-1} \qquad (19)$$

which also reduces to $S = \left[\sum_\ell \beta_\ell\right]^{-1}$ at the transformed means. In general estimates of β_ℓ will differ for the two functional forms (16) and (18) thus providing different estimates of aggregate returns to scale.

Local overall economies of scope for the translog model can be tested at the transformed means by computing the test statistics for cost complementarities (Fuss and Waverman (1977))

$$\frac{\partial^2 C}{\partial Q_k \, \partial Q_\ell} = \beta_k \beta_\ell + \delta_{k\ell} \qquad k \neq \ell \qquad (20)$$

The identical formula can be used to test for cost complementarities in the case of the hybrid translog function.

The translog cost function is undefined whenever one of the outputs is zero. As we have seen above, necessary and sufficient tests of economies of scope and tests of product-specific economies of scale require a cost function which is defined at zero levels of output. Fuss and Waverman (1980) circumvented that problem through their use of the hybrid translog function. They showed that for the hybrid function the degree of product j specific economies of scale could be computed at the transformed means as

$$S_j = \frac{\exp[\alpha_0] - \exp\left[\alpha_0 - \frac{\beta_j}{\theta} + \frac{\delta_{jj}}{2\theta^2}\right]}{\alpha_j \cdot \exp[\alpha_0]} \qquad (21)$$

Similarly $\tilde{S}_j$ can be computed as

$$\tilde{S}_j = \frac{\exp[\alpha_0] - \exp\left[\alpha_0 - \frac{\beta_j}{\theta} + \frac{\delta_{jj}}{2\theta^2}\right]}{\alpha_j \cdot \exp[\alpha_0]} \cdot Q_j^2 \qquad (22)$$

Finally, Fuss and Waverman (1980) showed that the degree of product specific economies of scope with respect to output 2 (private line services) for a 3 output hybrid translog cost function, calculated at the transformed mean, could be obtained as

$$SC_2 = \frac{\exp\left[\alpha_0 - \frac{\beta_2}{\theta} + \frac{\delta_{22}}{2\theta}\right] + \exp\left[\alpha_0 - \frac{1}{\theta}(\beta_1 + \beta_3) + \frac{1}{2\theta^2}(\delta_{11} + \delta_{33} + 2\delta_{13})\right] - \exp[\alpha_0]}{\exp[\alpha_0]} \qquad (23)$$

5.4 Evidence on Aggregate Economies of Scale

Table 4 presents a summary of the estimates of aggregate economies of scale, calculated at the mean of the sample (Denny et al (1979), Nadiri and Shankerman (1979), Christensen et al. (1980), Fuss and Waverman (1980)) or a mid-point observation (Fuss and Waverman (1977), Breslaw and Smith (1980)) along with approximate 95% confidence intervals where available.

The estimates for Bell Canada based on the translog cost function appear to indicate that the aggregate scale elasticity is in the neighborhood of 1.4. If this were true a 1% increase in all outputs would result in only a 0.7% increase in (long-run) total costs, a very substantial efficiency effect. The estimates of the scale elasticity for A.T.& T. based on the translog function are even higher. On the other hand, Fuss and Waverman's (1980) estimate based on the hybrid translog model is substantially below the other Canadian estimates and the U.S. estimates. Hence it is important to investigate the relationship between their estimate and the previous ones.

Table 1 demonstrates that the Fuss-Waverman (1980) structure differs from the previous Canadian studies in a number of ways related to data sets, output variable definition, technical change specifications, behavioural assumptions and estimation procedures. Yet from Table 4 it appears that the mean scale elasticity estimate for Bell Canada is invariant to these differences, as long as the translog cost function is used. Only when Fuss and Waverman switch to the hybrid translog specification does the scale elasticity change substantially. Essentially the decomposition of efficiency gains between scale effects and technological change effects is highly sensitive to variation in functional forms, even among second order flexible forms. This fact creates a real dilemma for policy decision-makers who wish

Table 4

Estimates of Aggregate Scale Economies for Telecommunications Production

	Fuss-Waverman (1977)	Fuss-Waverman (1977) revised*	Denny et al (1979)	Nadiri-Shankerman (1979)	Breslaw-Smith (1980)
Point Estimate	1.02	1.45	1.47	2.12	1.29
95% Confidence Region	(1.15, 0.89)		(1.59, 1.37)	(1.75, 2.69)	

	Denny-Fuss (1980)	Christensen et al (1980)**	Fuss-Waverman (1980) translog	Fuss-Waverman (1980) hybrid translog
Point Estimate	1.46	1.73	1.43	0.94
95% Confidence Region	(2.15, 1.10)	(1.94, 1.56)	(1.63, 1.26)	(1.09, 0.83)

* As reported in Denny et al (1979) using revised Bell Canada data.

** My best guess as to the preferred estimate - corresponds to Table 6, specification (10) based on Bell R&D Expenditures.

to base their decisions, at least in part, on those empirical estimates of scale economies which have been provided by the most current research.

Since the hybrid translog function contains the ordinary translog function as a nested special case traditional methods of statistical inference are available for discriminating among them. Recall that the hybrid translog function approaches the ordinary translog function as θ approaches 0. However at $\theta = 0$, the likelihood function becomes degenerate and hence this value cannot be imposed in estimating the hybrid function. Nevertheless the translog function can be approximated as closely as desired by choosing θ close to 0 . Fuss and Waverman chose $\theta = 0.01$. At that point, $\frac{Q_j^{*\theta} - 1}{\theta}$ is virtually identical to $\log Q_j^*$. The hybrid translog function with $\theta = 0.01$ and the ordinary translog function yield essentially identical empirical estimates. A likelihood ratio test of the null hypothesis $\theta = 0.01$ yielded the test statistic 11.92. The Chi-squared critical value is 3.84 (6.64) at the 5% (1%) significance level. At any reasonable significance level the null hypothesis was rejected, which implies rejection of the ordinary translog model and its associated estimates of substantial overall scale economies. Fuss and Waverman go on to conclude on the basis of these results that any aggregate economies of scale which exist are modest at best. This is the correct conclusion on the basis of formal statistical inference. However I think that the important lesson to be learned is that we still do not know the extent of aggregate scale economies in telecommunications despite the enormous amount of research effort devoted to that topic. The main value of Fuss and Waverman's research on this issue is to point out the danger of accepting for policy purposes at this time the evidence generated by the ordinary translog cost function estimates - that telecommunications production is subject to substantial increasing returns to scale in the aggregate. Guilkey and Lovell (1980) noted in their Monte Carlo study a tendency of the translog function to overestimate returns to scale. Perhaps this phenomon is at work here. In any case it is important to determine whether the Canadian and U.S. studies which use a single aggregate output suffer from the same lack of robustness to functional form specification as do the three output Canadian studies.

5.5 Evidence on Economies of Scope and Product-Specific Economies of Scale

In order to provide evidence on economies of scope and product-specific economies of scale one must estimate a multi-output technology. No U.S. studies have appeared as yet which disaggregate output, hence all evidence to date comes from Canadian studies. Fuss and Waverman (1977) using the local test (see equation (10)) find no statistically significant economies of scope. They do find insignificant cost complementarities between local and toll services and between toll and competitive services. Breslaw and Smith (1980) also using (10), found cost complementarities between local and toll services which were "unimportant relative to marginal cost". Fuss and Waverman (1980), using the global test outlined earlier found no statistically significant economies of scope between private line services and the other services. The evidence to date would appear to indicate that cost savings in telecommunications due to economies of scope are, at best, minor relative to aggregate costs.

The only study to investigate product-specific economies of scale was Fuss and Waverman (1980). Using equation (21) they estimated that Bell Canada produced private line services subject to increasing returns to scale.

However, from the estimation of equation (22) they determined that these returns to scale would be exhausted if Bell became the monopoly supplier of private line services. They concluded that there was no statistically significant static efficiency-related evidence that competition should not be encouraged in the provision of private line services.

6. Evidence on Technical Change

A number of the authors being surveyed have noted that the most difficult problem in estimating cost functionsfor telecommunications is the specification of technical change. In order to specify technical change one looks for an indicator of shifts in the cost function, i.e., a reason why costs might decline for a given set of factor prices and outputs. The most common technical change indicator used in econometric studies is the passage of time. This method has been used in telecommunications by Fuss and Waverman (1977) and Nadiri and Shankerman (1979). While time is simple to compute, it is itself a rather explicit indication of ignorance regarding the process of technical change. The research and development (R&D) expenditure pattern is another technical change indicator often used in econometric studies. Telecommunications studies which have used R&D effort include Nadiri and Shankerman (1979) and Christensen et al (1980). The main problem with this indicator is that it is a measure of input into the innovative process rather than output from the process. Outputs from the innovative process which have become embodied in telecommunications production include direct distance dialling facilities and improved (modern) switching facilities. These indicators of innovative activity have been used by Denny et al (1979), Denny and Fuss (1980), Breslaw and Smith (1980), Fuss and Waverman (1980) and Christensen et al (1980). They also played an important role in the decomposition analysis of Denny, Fuss and Waverman (1979a) where total factor productivity growth was decomposed into scale effects and effects due to technical change-inducing innovative activity. The innovations indicators used have been: (1) the percentage of telephones with access to direct distance dialling facilities (Denny et al, Denny, Fuss and Waverman, Denny and Fuss, Fuss and Waverman), (2) the percentage of long distance calls directly dialed (Christensen et al), (3) the percentage of telephones connected to central offices with modern switching facilities (Denny et al, Denny, Fuss and Waverman, Fuss and Waverman, Christensen et al) and (4), an index combination of (1) and (3) (Breslaw and Smith). These measures come closest to the spirit of technological change indicators but they suffer from the disadvantage that only a small number of major innovations are covered. Small-scale continuous technical change is not represented, nor is there any indicator which might represent improvements in outside plant (transmission facilities).

The technical change indicators have been incorporated into the specification of the cost function in a number of ways. First, the measure of technical chanqe can be treated as just another variable in the second order expansion (Nadiri and Shankerman (1979), Denny and Fuss (1980), Breslaw and Smith (1980)). Second, technical change can be specified as augmenting: capital augmenting (Fuss and Waverman (1977)), all factors augmenting (Christensen et al (1980)) and output augmenting (Denny et al (1979), Denny, Fuss and Waverman (1979a), Fuss and Waverman (1980)). At this point there does not appear to be evidence that any one method of incorporating technical change is the superior one.

In cost function models the rate of technical change is measured by the proportionate downward shift of the cost function over time.[8] For Bell Canada during the period 1952-76 this rate has been estimated at 0.8% (Denny et al (1979)). There exists no comparable estimate for A.T.& T. The direct technical change effect estimate given by Nadiri and Shankerman (1979) of 1.2% appears to be comparable but this appearance is misleading.

Nadiri and Shankerman did not permit the trend toward a higher scale elasticity over time contained in their estimated structure to affect their decomposition of total factor productivity growth into technical change effects and scale effects. Hence their estimated rate of technical change is inconsistent with their scale estimates and relative to the scale effect contains an upward bias. There is no evidence bearing on the question of whether technical change in telecommunications has been slower in Canada than in the United States.

The estimated rate of technical change can be used as an aid in the evaluation of scale elasticity estimates. Denny et al (1979) have shown that average cost can be decomposed into factor price effects, scale effects, and technical change effects by the formula

$$\overset{\cdot}{C/Q} = \sum_i S_i \dot{p}_i + (S^{-1} - 1)\dot{Q} + \dot{B} \qquad (24)$$

where C is cost; Q is output in the single output case and cost elasticity weighted aggregate output in the multi-output case; S is the overall scale elasticity; and $\dot{B}$ is technical change. The dot represents a rate of change. The rate of change of cost efficiency is given by $\overset{\cdot}{C/Q} - \sum S_i \dot{p}_i$ which can easily be calculated from time series data. The difficult problem is to decompose this cost efficiency into scale and technical change effects. The greater the scale elasticity S, the lower the rate of technical change $\dot{B}$.

One method of obtaining some perspective on the estimated scale and technical change effects decomposition is to compare the implied rate of technical change with that estimated for manufacturing industries. For the period 1963-76, Denny et al. (1979) estimated that $\dot{B} = 0.0064$ for Bell Canada. By way of contrast, Denny, Fuss and Waverman (1979b) estimated that the average rate of technical change for 20 two-digit industries over the period 1961-75 was $\dot{B} = 0.011$. Sixteen of the twenty industries had $\dot{B} > 0.0064$. One would not expect the rate of technical change in telecommunications to be only 60% of the manufacturing average. Nor would one expect the rate of technical change in telecommunications to be slower than that of 80% of the two-digit manufacturing industries. Hence Denny et al's $\dot{B}$ is probably a substantial underestimate, and therefore their $S = 1.47$ is likely to be a substantial overestimate of the true aggregate returns to scale; a fact which is consistent with the analysis of section 5.

Evidence relating to the bias in technical change among aggregate factors appears to be consistent in Canada and the United States. Denny et al (1979) and Nadiri and Shankerman (1979) found technical change to be capital using and labour saving. Denny et al found technical change to be materials saving whereas Nadiri and Shankerman found it to be materials neutral. With respect to specific technical change indicators, Denny et al found that the diffusion of direct distance dialing facilities through the Bell Canada telecommunications network was capital using and materials and labour saving. In contrast they estimated that the conversion to modern switching facilities resulted in savings with respect to all three factors of production.

Denny and Fuss (1980) also found technical change as represented by access to direct distance dialing facilities to be capital using and labour and materials saving. Among the occupational categories, the severity of the labour-saving impact was felt in inverse relation to the skill level associated with the occupation. Technical change had its strongest effect on the operators category, the category which one could expect to be the most directly influenced by the direct-distance dialing innovation.

7. Telecommunications Production and the Averch-Johnson Effect

Investor-owned telecommunications firms, such as Bell Canada and A.T.&T., are subject to rate of return regulation. It is well-known that rate of return regulation can bias the choice of inputs away from the cost-minimizing mix. This effect is known as the Averch-Johnson (A-J) effect. If the Averch-Johnson effect is present then parameters, and hence technological characteristics estimated from econometric cost functions, will be biased due to misspecification of the behavioural model. In this section we consider the way in which the A-J effect has been explicitly incorporated into econometric cost functions. Suppose the product transformation function is given by

$$F(Q_1 \ldots Q_m; K, X_2, \ldots X_n) = 0 \qquad (25)$$

where $K = X_1$ is the capital stock used to determine the allowed return. Then the firm's problem is to maximize:

$$\sum_{i=1}^{m} q_i Q_i - \sum_{j=2}^{n} p_j X_j - p_k \cdot K \qquad (26)$$

subject to (25) and

$$\sum q_i Q_i - \sum p_j X_j \leq sK \qquad (27)$$

where q_i, $i=1,\ldots m$ are endogenous output prices, p_k is the price of capital services and s is the allowed rate of return.

Fuss and Waverman (1977) showed that one solution to the above constrained maximization problem can be formulated as follows. Assuming an ex ante binding rate of return constraint ((27) holds with equality), the cost function takes the form

$$C = C(p_k, p_2, \ldots p_n, s, Q_1 \ldots Q_m) \quad . \tag{28}$$

As shown by Fuss and Waverman (1977), the system to be estimated consists of the cost function (28), the marginal profitability conditions $MR_i = MC_i$, where MR_i is marginal revenue and MC_i is marginal cost, and the input demand functions generated by the modified Shephard's Lemma:

$$\frac{\partial C}{\partial p_j} = X_j(1-\lambda_1) \qquad j = 2,\ldots n$$

$$\frac{\partial C}{\partial p_k} = K \tag{29}$$

$$\frac{\partial C}{\partial s} = -\lambda_1 K$$

λ_1 is the Lagrangian multiplier associated with constraint (27).

Actual estimating equations can be formed by noting that

$$X_j/X_\ell = \frac{\partial C}{\partial p_j} \Big/ \frac{\partial C}{\partial p_\ell} \tag{30}$$

which eliminates the unknown Lagrangian multiplier λ_1. This multiplier can be obtained from the above equations as

$$\lambda_1 = -\frac{\partial C}{\partial s} \Big/ \frac{\partial C}{\partial p_k} \tag{31}$$

An alternative rate of return constraint model has been considered by Diewert (1979) and Fuss and Waverman (1980). This model centres around the variable (short run) cost function. Suppose that the regulated firm minimizes the cost of producing the vector of outputs $Q = \{Q_1, \ldots Q_m\}$ conditional on the capital stock at the beginning of the period (K_{-1}). In that case there exists a variable cost function.

$$VC = VC(p, K_{-1}, Q) \tag{32}$$

(where $p = (p_2, p_3, \ldots p_n)$)

with the following properties:

(33)

(a) VC is concave in p

(b) VC is linear homogeneous in p

(c) VC is increasing in p, Q and decreasing in K_{-1} (monotonicity)

(d) $\frac{\partial VC}{\partial p_j} = X_j$ (Shephard's Lemma)

The conditional profit maximizing problem for the rate of return regulated utility can now be written as: choose outputs Q_i and output prices q_i so as to maximize:

$$\text{Profit} = \sum_i q_i Q_i - VC(p, K_{-1}, Q) - p_k \cdot K_{-1} \tag{34}$$

subject to

$$\sum q_i Q_i - VC(p, K_{-1}, Q) = sK_{-1} \tag{35}$$

The solution to the constrained maximization problem can be shown to be (Diewert (1979), Fuss and Waverman (1980)).

$$-\frac{\partial VC}{\partial K} = \frac{p_k - \mu s}{1 - \mu} = p_k^* \tag{36}$$

where μ is the Lagrangian multiplier associated with constraint (35).

The right hand side of (36) is the shadow price of capital p_k^*. The left hand side can be computed once the parameters of VC are known. It can be shown that (36) can be solved for μ in the form

$$\mu = \frac{M_1 + \frac{\partial \log VC}{\partial \log K}}{M_2 + \frac{\partial \log VC}{\partial \log K}} \tag{37}$$

where $M_1 = \frac{p_k \cdot K}{VC}$, $M_2 = \frac{sK}{VC}$.

Fuss and Waverman (1980) have shown that the parameters of this model can be estimated from a system consisting of the cost function (32), the input demand functions obtained from (33d) and equations derived from the marginal profitability conditions $MR_i = MC_i$.

The two models of a regulated utility outlined above were estimated by Fuss and Waverman (1980) for Bell Canada. They concluded that neither model was supported by the data. Similar negative results concerning the A-J effect have been reported by Breslaw and Smith (1980). It would appear that if Bell Canada is typical of investor-owned telecommunications firms, input use inefficiency due to rate of return regulation is, at worst, a minor problem.

8. A Final Overview

Recent empirical studies of telecommunications production have utilized the theory of duality between cost and production and the availability of flexible functional forms to allow for the possibility of general substitution effects, non-homothetic scale effects, and non-neutral technical change effects. An important result of this research activity has been a substantial advance in the level of methodology applied in the telecommunications area. Much has been learned about substitution possibilities and the nature of technical change. Output expansion effects bearing on the natural monopoly question remain controversial. Perhaps this is inevitable, given the highly trended output and technical change indicator data in the typical time series data set which is available in telecommunications. This trending makes the separation of efficiency gains into those due to scale economies and those due to technical change very difficult. Because of this difficulty any researcher trying to establish the extent of scale economies in telecommunications from time series data should report as much detail on technical change estimates as on scale estimates. It is only relative to the reasonableness of technical change estimates that the appropriateness of scale estimates can be evaluated. This fact has not often been appreciated by those doing research in this area, including the present author.

Reliable estimates of economies of scale and economies of scope will probably have to await the development of a pooled time series-cross section data base. Such a data base might be formed from the operating companies of the U.S. Bell System or several Canadian telephone companies. In this latter regard, I note with approval the recent efforts of a number of Canadian telephone companies, in cooperation with the Department of Communications, to begin the construction of the needed data.

Footnotes

1. The cost function was first applied to telecommunications by Waverman (1976) who utilized a Cobb-Douglas cost function. Fuss and Waverman (1977) applied the multiple output cost function to telecommunications and were the first to exploit the empirical implications of duality theory in telecommunications applications.
2. Virtually all regulatory problems in telecommunications are linked to characteristics of the cost structure. Prominent examples are the economies of scale - economies of scope natural monopoly debate featured in the recent CNCP-Bell interconnection case, and the discussion concerning cross-subsidization of basic local service found in recent Bell Canada rate increase application hearings.
3. Telecommunications firms compete in local and national markets for labour and material inputs. Most telecommunication firms purchase equipment in international markets. Bell Canada and A.T.&T. purchase virtually all their equipment requirements from their subsidiaries, (Northern Telecom. and Western Electric respectively) and may engage in artificial transfer pricing. However, detailed regulatory scrutiny of these equipment purchases probably results in purchase prices which reflect international competitive conditions.
4. Fuss and Waverman (1977) and Diewert (1979) developed cost function models which incorporate the rate of return constraint. Fuss and Waverman (1980) attempted to estimate these models for Bell Canada but were unsuccessful. This lack of success suggests that the Averch-Johnson effect may be unimportant in Canadian telecommunications.

5. Denny and Fuss (1980) employ a two-stage translog specification for which the calculation of the price elasticities are more complex than equations (5) and (6). The reader is referred to their article for details.

6. Baumol (1977) calls these output charac:eristics "decreasing ray average cost" and "transray convexity" respectively.

7. This function applies a Box-Cox transformation to outputs. The form used here was first proposed by Caves, Christensen and Tretheway (1980). The use of the Box-Cox transformation has a long history in econometrics. For a recent example of its use in the context of a single output cost function, see Berndt and Khaled (1979).

8. For the link between the downward shift of the cost function and the upward shift of the production function as measures of technical change see Denny, Fuss and Waverman (1979).

ECONOMIES OF SCALE AND SCOPE IN BELL CANADA:

SOME ECONOMETRIC EVIDENCE

F. KISS

S. KARABADJIAN

B. LEFEBVRE

Bell Canada

1. INTRODUCTION

The main objective of this paper is to establish, by means of aggregate econometric modeling of Bell Canada's production structure, whether there are internal economies inherent in the process of producing telecommunication services by the company in large quantities (economies of scale) and great variety (economies of scope).

It is assumed that the services of Bell Canada's productive factors (labour, capital, etc.) can be related to the output (telecommunication service) volumes the company produces by a transformation function, which exhibits certain useful and economically meaningful mathematical properties (continuous twice differentiable, strictly monotone and quasi-concave). Further, it is assumed that Bell's production technology can be expressed equivalently by a dual cost function, relating exogenous output levels and input prices to the company's total production cost. Technological changes are regarded as shifts in the transformation and cost functions.

The statistical estimation of cost functions is pursued in the paper. The translog (TL) and a generalized form of the translog (GTL) cost function are chosen from the several highly general flexible functional forms that were introduced in the literature during the 1970's.

The estimation effort is guided by deductive reasoning. A very elaborate form of the GTL cost function is estimated first. Then, statistical tests on restrictive hypotheses and an analysis of the estimated parameters and economic properties are used to gradually restrict the specification, thereby reducing its degree of generality and making it more reflective of the specific economic characteristics of Bell Canada's technology. This process is pursued until the limits of statistically justifiable and economically meaningful restrictions are approached.

A comparison of the estimation results obtained from the gradually restricted cost functions serves the second objective of the paper, which is the examination of the robustness of empirical findings on internal economies. To further such an examination, the sensitivity of internal economies to sample variation and alternative variable measurements is also observed in models which appear to be preferable to other estimated models.

The structure of the paper is as follows. The applied cost functions are described, several forms of internal economies are defined and the statistical testing procedures are established in Section 2. Section 3 contains a description of empirical results. Sub-sections are devoted to 3-output, 2-output and single output models. Section 4 offers a summary of conclusions, with some references to the evidence[1] generated by other econometric studies of Bell Canada.

2. COST FUNCTIONS, INTERNAL ECONOMIES, STATISTICAL TESTS

Increasing exploration of the theory of duality during the 1970's led to the recognition that cost functions were more suitable than production functions for estimating the characteristics of the Bell Canada production process.[2] The neoclassical cost function can be written as

$$(1) \quad C = g(Q_1, \ldots, Q_n; W_1, \ldots, W_m; T) \quad ,$$

where $Q_1, \ldots, Q_n$ denote output volumes, $W_1, \ldots, W_m$ denote factor input prices, T is an index of technological change and C refers to the total production cost, defined as the sum of payments to m factor inputs $X_1, \ldots, X_m$, i.e., $C = \sum_i W_i X_i$.

The translog (TL) and generalized translog (GTL) forms of the cost function are used to represent Bell Canada's production structure.[3] Each is a class of the flexible second order Taylor series approximation to the general form in equation (1). Both functional forms contain the natural logarithms of $W_1, \ldots, W_m$ and C, but GTL substitutes the Box-Cox transformation for the natural logarithms of $Q_1, \ldots, Q_n$ and/or T in the translog form.[4] The Box-Cox transformation of output Q_k is $(Q_k^{\lambda_k}-1)/\lambda_k$ $(k = 1, \ldots, n; \lambda_k \neq 0)$ and the technology variable is transformed in a similar fashion as $(T^{\lambda_T}-1)/\lambda_T$ $(\lambda_T \neq 0)$. The Box-Cox expression reduces to the logarithmic transformation of Q_k and T if the respective λ_k and λ_T values are zero. Hence, the GTL specification contains TL as a special limiting case.

To simplify the presentation of the models, equation (2) below contains Q_k^* and T^* variables, which represent both the logarithmic and the Box-Cox transformations; i.e., $Q_k^* = \log Q_k$ and $T^* = \log T$ in TL and $Q_k^* = (Q_k^{\lambda_k}-1)/\lambda_k$ and $T^* = (T^{\lambda_T}-1)/\lambda_T$ in GTL.

The TL/GTL cost function can now be written as

$$\begin{aligned}(2)\quad \log C = {} & \alpha_0 + \sum_{i=1}^{m} \alpha_i \log W_i + \sum_{k=1}^{n} \alpha_{Qk} Q_k^* + \beta T^* \\ & + (\tfrac{1}{2}) \sum_{i=1}^{m} \sum_{j=1}^{m} \gamma_{ij} \log W_i \log W_j + (\tfrac{1}{2}) \sum_{k=1}^{n} \sum_{l=1}^{n} \delta_{kl} Q_k^* Q_l^* \\ & + \sum_{i=1}^{m} \sum_{k=1}^{n} \rho_{ik} \log W_i Q_k^* + \sum_{i=1}^{m} \beta_i \log W_i T^* \\ & + \sum_{k=1}^{n} \beta_{Qk} Q_k^* T^* + (\tfrac{1}{2}) \beta_T (T^*)^2\end{aligned}$$

where the variables are defined as in equation (1) above.[5]

The cost function is constrained to be homogeneous of degree one in the input prices[6] by the following set of restrictions:

$$(3)\quad \sum_{i=1}^{m} \alpha_i = 1; \quad \sum_{i=1}^{m} \gamma_{ij} = \sum_{i=1}^{m} \rho_{ik} = \sum_{i=1}^{m} \beta_i = 0 \quad (j=1,\dots,m;\ k=1,\dots,n).$$

The symmetry conditions in a second order approximation together with (3) imply that

$$(4)\quad \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0 .$$

Since the number of parameters to be estimated is usually large, it is advisable to use additional information to construct a more complete model of the cost structure. Assuming that the cost minimizing input levels are chosen to produce the observed output volumes, invoking a lemma ($\partial C/\partial W_i = X_i$) by Shephard (1970) and partially differentiating the cost function with respect to input prices, cost share equations for each input are constructed as

$$(5)\quad S_i = \alpha_i + \sum_{j=1}^{m} \gamma_{ij} \log W_j + \sum_{k=1}^{n} \rho_{ik} Q_k^* + \beta_i T^* \qquad (i=1,\ldots,m).$$

Equations (2) and (5) are estimated simultaneously. Since the parameters of (5) are a subset of those of (2), the cost share equations increase the available degrees of freedom and improve statistical precision. Following Christensen and Greene (1976), disturbance terms are added to each cost share equation to reflect random errors in optimization. Since the cost shares sum to unity, their disturbances sum to zero. To preserve the non-singularity of the covariance matrix, one of the m cost share equations[7] is deleted from the estimation process. Maximum likelihood parameter estimates are invariant to the deleted equation. Using the iterative estimation technique for seemingly unrelated equations of Zellner (1962) on a large sample of Bell Canada data ensures that maximum likelihood estimates are obtained if the covariance matrix converges.

Bell Canada is assumed to be a cost minimizer and all the n outputs in equations (2) and (5) are exogenous. This assumption conforms with the single output models of Bell Canada by Denny et al (1979) and Smith and Corbo (1979), and also with single output models of the Bell System by Nadiri and Shankerman (1979) and Christensen et al (1980), but differs from multi-output specifications by Fuss and Waverman (1977, 1978, 1980), Smith and Corbo (1979) and Denny et al (1979), where the local service output is exogenous but the toll service output is assumed to be endogenously determined by Bell Canada (through endogenously set prices), resulting from monopolistic profit maximizing behaviour (marginal cost equals marginal revenue)[8].

Four distinct forms of internal economies are defined below. These are (1) overall economies of scale, (2) cost complementarity between outputs, (3) global economies of scope and (4) output-specific economies of scale.

Overall economies of scale are measured by the inverse of the sum of cost elasticities with respect to outputs. This statistic, denoted by ε and called scale elasticity[9] below, can be derived from the general cost function as

$$(6) \quad \varepsilon = \left(\sum_{k=1}^{n} \frac{\partial \log C}{\partial \log Q_k} \right)^{-1} .$$

The TL form of equation (2) yields the following expression

$$(7a) \quad \varepsilon = \left[\sum_{k=1}^{n} (\alpha_{Qk} + \sum_{i=1}^{m} \rho_{ik} \log W_i + \sum_{l=1}^{n} \delta_{kl} \log Q_l + \beta_{Qk} \log T) \right]^{-1}$$

and the scale elasticity is derived from the GTL function as

$$(7b)\quad \varepsilon = \left[\sum_{k=1}^{n} \{ Q_k^{\lambda_k} [\alpha_{Qk} + \sum_{i=1}^{m} \rho_{ik} \log W_i + \sum_{l=1}^{n} \delta_{kl} \frac{Q_l^{\lambda_l} - 1}{\lambda_l} + \beta_{Qk} \frac{T^{\lambda_T} - 1}{\lambda_T}] \} \right]^{-1} .$$

Local overall economies (diseconomies) of scale are said to exist if $\varepsilon > 1$ ($\varepsilon < 1$), while the underlying technology is characterized locally by neither economies nor diseconomies of scale if $\varepsilon=1$. The latter is also referred to as constant returns to scale.

At the expansion point, where $W_i = Q_k = T = 1$, the cost elasticity with respect to Q_k reduces to α_{Qk} in both functional forms; hence the scale elasticity is

$$(8)\quad \varepsilon = (\sum_{k=1}^{n} \alpha_{Qk})^{-1} .$$

The hypothesis of constant returns to scale can be tested by constructing confidence limits for the terms on the right hand side of equations (7a) and (7b) and observing if the value of $\varepsilon=1$ falls within or outside the confidence interval. The procedure can be simplified if the confidence limits are computed for the expansion point only, using equation (8).

Likelihood ratio tests, which are used extensively to test the validity of various assumptions resulting in parametric restrictions, are also utilized in the process of testing for the hypothesis of constant returns to scale. The following parametric restrictions are imposed on the TL/GTL[10] equation by hypothesizing constant returns to scale:

$$(9)\quad \sum_{k=1}^{n} \alpha_{Qk} = 1; \sum_{k=1}^{n} \rho_{ik} = \sum_{l=1}^{n} \delta_{kl} = \sum_{k=1}^{n} \beta_{Qk} = 0 \quad (i = 1,\dots,m).$$

The number of restrictions in a given specification depends on the number of outputs (n) and inputs (m). Since the parameters are maximum likelihood estimates, the log of the likelihood function for the estimates with restricted and unrestricted parameters (Ω_R and Ω_U, respectively) can be used in likelihood ratios of the form $\lambda = \Omega_R - \Omega_U$. -2λ is distributed asymptotically as Chi-squared, with degrees of freedom equal to the number of independently imposed restrictions on the parameters, if the restrictive null hypothesis is true. Normally, the comparison is made between the computed -2λ and the χ^2 value at the .05 level; however, in some cases the χ^2 value at the .01 level is also considered. The null hypothesis cannot be rejected (is rejected) if the critical χ^2 value is less (greater) than the computed -2λ.

Cost complementarity gives a local estimate of economies of scope at specific output levels. The test statistic for cost complementarity in a twice differentiable cost function is the second order cross-derivative of the cost function with respect to any two outputs:[11]

$$(10)\quad CC_{kl} = \frac{\partial^2 C}{\partial Q_k \, \partial Q_l} \qquad (k \neq l;\ k,l = 1,\dots,n) \quad .$$

Cost complementarity exists, when CC<0; i.e., when infinitesimal increases/decreases in the volume of one output make the marginal cost of the other output decline/increase.

In the translog model, the test statistic can be written as

$$(11a)\quad CC_{kl} = \frac{C}{Q_k \, Q_l} \left[\frac{\partial \log C}{\partial \log Q_k} \cdot \frac{\partial \log C}{\partial \log Q_l} + \delta_{kl} \right]$$

and in the GTL model it becomes

$$(11b)\quad CC_{kl} = \frac{C}{Q_k\, Q_l}\left[\frac{\partial \log C}{\partial \log Q_k} \cdot \frac{\partial \log C}{\partial \log Q_l}\right] + \delta_{kl} \cdot C \cdot Q_k^{\lambda_k - 1} \cdot Q_l^{\lambda_l - 1} \quad .$$

At the expansion point, the cost complementarity test statistic in both models reduces to

$$(12)\quad CC_{kl} = \alpha_{Qk}\, \alpha_{Ql} + \delta_{kl} \quad .$$

The null hypothesis of no cost complementarity can be tested by constructing confidence intervals for terms on the right hand side of equations (11a) or (11b), or equation (12) when the test is performed at the expansion point, and observing if the value CC=0 falls within or outside the confidence interval.

<u>Economies of scope</u> exist globally (in the entire range of output volumes) when the joint production of an industy's outputs is cheaper than their separate production; i.e., when

$$(13)\quad C(Q_1,\ldots,Q_n) < C(Q_1,0,\ldots) + C(0,Q_2,0,\ldots) + \ldots + C(0,\ldots,0,Q_n),$$
$$(Q_1,\ldots,Q_n) > 0 \quad .$$

This case is referred to as that of <u>overall</u> economies of scope. Output-specific economies of scope exist, when the joint production of an output (Q_k) with the existing combination of other outputs is cheaper than its separate production; i.e., when

$$(14)\quad C(Q_1,\ldots, Q_n) < C(Q_1,\ldots,Q_{k-1},0,Q_{k+1},\ldots,Q_n) + C(0,\ldots,0,Q_k,0,\ldots),$$
$$(Q\ ,\ldots,Q_n) > 0 \quad .$$

Following Panzar and Willig (1979), a test statistic for Q_k-specific economies of scope can be written as

$$(15)\quad SCOPE_k = \frac{C(Q_1,\ldots,Q_{k-1},0,Q_{k+1},\ldots,Q_n) + C(0,\ldots,0,Q_k,0,\ldots) - C(Q_1,\ldots,Q_n)}{C(Q_1,\ldots,Q_n)} \quad .$$

Q_k-specific economies of scope exist when $SCOPE_k > 0$. To simplify the procedure, tests of output-specific economies of scope are carried out at the expansion point only. The translog cost function is not well defined for zero output levels; thus, it is not suited to carry out tests of global economies of scope. The economies of scope statistic can be derived from the GTL function as

(16)
$$SCOPE_k = \frac{\exp[\ \alpha_0 - \frac{\alpha_{Qk}}{\lambda_k} + \frac{1}{2}\frac{\delta_{kk}}{\lambda_k^2}\ ] + \exp[\ \alpha_0 - \sum_{\substack{i=1\\ i\neq k}}^{n} \frac{\alpha_{Qi}}{\lambda_i} + \frac{1}{2} \sum_{\substack{i=1\\ i\neq k}}^{n} \sum_{\substack{j=1\\ j\neq k}}^{n} \frac{\delta_{ij}}{\lambda_i\ \lambda_j}\ ] - \exp[\alpha_0]}{\exp[\alpha_0]} .$$

The null-hypothesis of no economies of scope is tested by constructing confidence intervals for the terms on the right hand side of equation (16) and observing whether the value $SCOPE_k = 0$ falls within or outside the confidence limits.

Output-specific economies of scale in a multi-product firm result from less than proportional increases/decreases in the cost specific to an output, when the level of that output increases/decreases, while all other output levels remain unchanged. Q_k-specific cost is the addition to the total cost of production that results from Q_k being produced. It is the incremental cost of $Q_k(IC_k)$:

$$IC_k \ = \ C(Q_1,\ldots,Q_n) - C(Q_1,\ldots,Q_{k-1},0,Q_{k+1},\ldots,Q_n)$$

The average incremental cost of Q_k is defined as $AIC_k=IC_k/Q_k$. AIC_k declines if Q_k-specific economies of scale are present. In this case, AIC_k is greater than the marginal cost of Q_k and their ratio is greater than one.

Following Panzar and Willig (1979), the degree of Q_k-specific economies of scale is defined as the ratio between the average incremental cost and marginal cost of Q_k and can be expressed as

$$(17)\quad S_k = \frac{IC_k/Q_k}{\partial C/\partial Q_k} = \frac{IC_k/C}{\varepsilon_{CQk}} \quad ,$$

where $C = C(Q_1,\ldots,Q_n)$ and $\varepsilon_{CQk} = \partial \log C/\partial \log Q_k$. There are economies of scale specific to Q_k if $S_k > 1$.

The translog cost function is not suitable for the determination of IC_k, because it is not well defined for zero output levels. Fuss and Waverman (1980) have shown that the degree of output-specific economies of scale can be estimated with relative ease at the expansion point, where $W_i = Q_k = T = 1$. The GTL function yields the following expression for Q_k-specific economies of scale:

$$(18)\quad S_k = \frac{\exp[\alpha_0] - \exp[\alpha_0 - \alpha_{Qk}/\lambda_k + \delta_{kk}/2\lambda_k^2]}{\alpha_{Qk} \cdot \exp[\alpha_0]} \qquad (k=1,\ldots,n) \quad .$$

Testing for the null-hypothesis of no Q_k-specific economies of scale is done again by constructing confidence intervals for the terms on the right hand side of equation (18) and by observing if $S_k=1$ falls within or outside the confidence interval.

Panzar and Willig (1979) showed that overall and output-specific economies of scale and economies of scope are associated by the following relationship

$$(19)\quad \varepsilon = \frac{w_k S_k + (1 - w_k)S_\eta}{1\text{-}SCOPE_k} \qquad \eta = \{1,\ldots,k\text{-}1,k\text{+}1,\ldots,n\}$$

where

$$w_k = Q_k \frac{\partial C}{\partial Q_k} \Big/ \sum_{i=1}^{n} Q_i \frac{\partial C}{\partial Q_i} \geq 0 \quad \text{and} \quad S_\eta = IC_\eta \Big/ \sum_{\substack{i=1 \\ i\neq k}}^{n} Q_i \frac{\partial C}{\partial Q_i}$$

ε, $SCOPE_k$ and S_k are as defined in equations (6), (15) and (17) respectively. IC_η is the incremental cost specific to the product set η; i.e.,

$$IC_\eta = C(Q_1,\ldots,Q_n) - C(0,\ldots,Q_k,\ldots,0).$$

In the absence of economies of scope, ε will reduce to the weighted average output-specific returns to scale. However, when economies of scope are present, the multiproduct technology will exhibit greater overall returns to scale than the weighted average of the output-specific scale elasticities.

3. EMPIRICAL RESULTS

3.1 Models with Three Outputs

Considering that a high degree of multicollinearity may be introduced in the TL/GTL cost model by the large number of second order terms, it was decided that the estimation of the cost function would not be attempted with more than three output and three input variables. The three output volumes are represented by Törnqvist volume indices of Bell Canada's local, directory advertising and miscellaneous service outputs (Q_1), message toll (intra-Bell, Canada, US and overseas) and WATS outputs (Q_2) and private line, TWX and other toll outputs (Q_3). Q_1 is simply referred to as local output, Q_2 as message toll output and Q_3 is called other toll output. The three input price variables are those of labour (W_1), capital (W_2) and material (W_3). Labour price is the implicit Törnqvist price index of labour, computed as the ratio of the index of total labour cost to the Törnqvist volume index of labour. Capital price is a measure of the user cost or rental price of capital and material price is the implicit Törnqvist price index of materials, rents, supplies and services. The sample on which the models were estimated contains annual data for the period 1952 to 1978. The variables are normalized around their respective 1967 values. The data are described in detail in Kiss (1981). The index of technological changes is described and shown in Appendix C. The T2 variable in Table C.1 was used in all models.

The most general and elaborate presentation of the underlying technology is the unconstrained 3-output 3-input GTL function, which was estimated first. Although most parameters were significant and the model happened to fit the data very well, many of the estimated economic properties of the function were unacceptable. The majority of the annual estimates of the marginal cost of other toll were negative, the cost elasticity with respect to technological changes was positive in seven years and its values were unreasonably low at the end of the observation period. The curvature was generally incorrect. Scale elasticity was estimated at 1.43 at the expansion point (1967). Several unacceptably high values appeared among the annual estimates. The estimated CC values indicated cost complementarity between local and message toll and local and other toll services during the entire sample period. However, the estimated CC_{13} values were unrealistically low for the first few years of the period. Global economies of scope and output-specific economies of scale statistics could not be computed, because two of the three λ_k values were negative. The model appeared to be too general to be successfully estimated with the available amount of information. Since additional information did not offer itself (profit maximizing behaviour was a priori rejected for Bell Canada) and the results did not suggest simplifying parameter restrictions such as homogeneity, homotheticity or input-neutral technological changes, further experimentation was carried out by simplifying the Box-Cox transformation, even though these simplifications were contrary to the results of statistical tests.

It was assumed that $\lambda_Q=\lambda_1=\lambda_2=\lambda_3$ and a constrained model with a single λ_Q was estimated. The estimation results did not improve and although many of the estimated economic properties underwent significant numerical changes, the same problems were exhibited as in the unconstrained model (negative marginal costs, meaningless output-specific economies of scale and scope statistics). Only the curvature improved by becoming concave at 85% of the data points. Economies of scale were indicated at the expansion point ($\varepsilon=1.26$) and for the second half of the observation period, especially for the 1970's ($\varepsilon=1.52$-1.96), but the annual scale elasticity estimates did not compose a realistic pattern (declines during the years of introduction of crossbar and DDD technologies and high year-to-year fluctuation). Cost complementarity was indicated in each year between local and message toll and local and other toll, but not between message and other toll, services.

The non-linearity of the function with respect to the λ_k parameters was eliminated by logarithmically transforming the output variables (λ_k=0,k=1,2,3). The results were unacceptable as the concavity requirement was violated at all observation points.

Since the λ_T estimates were insignificant in both constrained 3-output GTL models, the transformation of the technology variable was restricted to be logarithmic (λ_T=0). The model was estimated with and without the $\lambda_1=\lambda_2=\lambda_3$ constraint. The curvature of the cost function was incorrect, when three different λ_k parameters were estimated for the outputs, but the concavity conditions were met at most observations in the single λ_Q model. (Nevertheless, the three λ_k's were significantly different from each other.) In the single λ_Q model, marginal costs were negative in several years for each of the three outputs and the output-specific economies of scale and scope statistics remained meaningless. Economies of scale were indicated (e.g., ε=1.22 at the expansion point and ε=1.43-1.76 during the 1970's) and the extremely large annual scale elasticity estimates that were obtained at the beginning of the sample period in the $\lambda_1\neq\lambda_2\neq\lambda_3$ model disappeared. Cost complementarity was indicated between local and other toll services in each year of the sample, but only the last three years showed complementarity between local and message toll services. As in the case of the full GTL model, parametric restrictions (homotheticity, homogeneity or input neutrality of technological changes) were not suggested by the results.

The 3-output 3-input TL model was also estimated. The concavity conditions were not met in the translog cost function.

In summary, the 3-output TL/GTL model proved to be too general. Many estimated parameters exhibited a high degree of instability and close to 50% of them were insignificant in the constrained models. As a result of instability in the parameters that appear in the test statistics, the estimation of cost complementarity and output-specific economies of scale and scope was not successful.[12] However, the scale elasticities were only slightly influenced by estimation problems. The annual estimates show a fairly realistic pattern in the two well-behaved cost functions and the following expansion point values (with asymptotic standard errors in parentheses) give a uniform indication of overall economies of scale:

Model	ε
1. Unconstrained GTL	1.43* (.09)
2. $\lambda_Q=\lambda_1=\lambda_2=\lambda_3$	1.26 (.15)
3. $\lambda_k=0$ (k=1,2,3)	1.67* (.18)
4. $\lambda_T=0$	1.43* (.10)
5. $\lambda_Q=\lambda_1=\lambda_2=\lambda_3$; $\lambda_T=0$	1.22 (.12)
6. TL ($\lambda_k=\lambda_T=0$; k=1,2,3)	1.67* (.18)

*Significantly greater than one at the .05 level.

3.2 Models with Two Outputs

With the failure of the 3-output TL cost function to produce realistic estimates of the economic properties of Bell Canada's technology, the possibilities offered by the 3-output model were exhausted and the number of outputs was reduced to two by aggregating message toll and other toll services. The output-related parameters were generally poorly estimated in the 3-output model (instability, insignificance) and it was hoped that their quality could be improved by reducing their number. There is additional justification for the toll aggregate (Q_2) in the fact that the data (more specifically the price indices) on other toll services are of considerably lower quality than on other outputs. The aggregation of the two toll categories minimizes the consequences of data problems, because other toll services represent relatively small volumes (12 to 15% of total constant dollar toll revenue during the last 15 years).

The unconstrained 2-output GTL cost function contains λ_1, λ_2 and λ_T parameters. A detailed account of the estimation results, together with those of constrained models, is given in Appendix A. The reduction in the number of outputs resulted in a marked improvement of the estimation results over those of the 3-output GTL and TL models. The concavity conditions were met at most observations, cost elasticities and marginal costs emerged with the a priori correct sign for all years. However, the remaining problems proved to be still rather serious. λ_1 (related to local output) acquired a negative sign; thus, the economies of scope and output-specific economies of scale tests that require cost estimates for zero output levels became meaningless. The annual estimates of cost elasticity with respect to technology showed a sharp decline between 1971 and 1978 (from -.41 to -4.1) and the scale elasticities were correspondingly reduced from 1.68 to .63. The insignificant estimates of all ρ_{ik} parameters (see Table A.3) suggested that the underlying technology might be homothetic. The homothetic function could not be estimated, because the covariance matrix did not converge. Likelihood ratio tests were carried out on various restrictions of the Box-Cox parameters (λ_k, λ_T). Even though all λ parameters were significant in the full model and all hypotheses (restrictions) were rejected (see Table A.5), further experiments were conducted by simplifying the transformation of outputs and technology in the same manner as in the case of the 3-output models.

The results improved when the function was estimated under the $\lambda_1=\lambda_2$ constraint. The decline in the technology elasticity of cost at the end of the period was more moderate than in the unconstrained model. An additional problem was created by the marginal cost of local services, which became negative in the last three years. However, when the model was further restricted by the $\lambda_k=0$ constraint (logarithmic output variables), all marginal costs were positive and the technology elasticity of cost, and consequently the scale elasticity estimates, showed further improvement.[13] Both the t-test and the likelihood ratio test suggested that $\lambda_T=0$; thus, the technology variable should appear in logarithmic form.

The results did not improve when the transformation of the technology variable was restricted to be logarithmic, and additional difficulties (some negative marginal costs for toll services) were encountered when the constraint $\lambda_1=\lambda_2$ was applied. The inexplicable sharp declines in the technology elasticity of cost reappeared at the end of the period. It seemed that this problem could be lessened only by the logarithmic transformation of the output variables; therefore, the 2-output translog cost function was also estimated, even though a likelihood ratio test rejected the TL in favour of the GTL model.

Although the translog model produced a slightly worse fit than the GTL models, it brought about further improvements in the estimated economic properties. Marginal costs had the a priori correct sign and the estimates of cost elasticity with respect to technological changes became more realistic. The still remaining estimation problems were related to factor substitution (insignificant partial elasticity of substitution between labour and capital) and the price elasticity of factor demand (incorrect sign for capital during the first six years of the period).

The evidence on overall economies of scale that emerges from the 2-output TL/GTL cost models is remarkably robust. All models produced significant estimates of economies of scale at the expansion point and the estimated scale elasticities fell into a relatively narrow range between 1.44 and 1.66. The annual scale elasticity estimates exhibited a pattern which had realistic features (lower values in the early years and an increase in the degree of scale economies, resulting from the introduction of crossbar and DDD technologies). A marked decline in the technology elasticity of cost after 1970 in all models seems to suggest that the scale elasticities of the 1970's were underestimated to varying degrees. However, this problem led to radical consequences only in the case of the unconstrained model. In the constrained models, the estimated scale elasticities range between 1.18 and 1.49 during the last three years of the observation period.

The following scale elasticity estimates (with their asymptotic standard errors in parentheses) were obtained from the 2-output GTL and TL models at the point of expansion:

Model	ε*	
1. Unconstrained GTL	1.44	(0.06)
2. $\lambda_Q=\lambda_1=\lambda_2$	1.50	(0.15)
3. $\lambda_k=0(k=1,2)$	1.64	(0.14)
4. $\lambda_T=0$	1.44	(0.11)
5. $\lambda_Q=\lambda_1=\lambda_2$; $\lambda_T=0$	1.44	(0.15)
6. TL ($\lambda_k=\lambda_T=0$; k=1,2)	1.66	(0.14)

*All estimates are significantly greater than one at the .05 level.

The 2-output models generated a uniform, but rather weak, indication of cost complementarity between Bell Canada's local and toll service outputs. Uniformity is indicated by the fact that the estimated cost complementarity (CC) statistics were negative in each year of the sample period in four out of six models, while one model yielded negative values for 78% of the observations. Only the unconstrained GTL model produced mostly positive CC statistics. The weakness of the estimates lies partly in their statistical insignificance at the expansion point and partly in the fact that five out of six models produced upward or downward trended estimates of CC. The indication of global economies of scope was unanimous, but weak. Each of the three constrained GTL models in which the estimation could be meaningfully accomplished generated positive values for SCOPE in each year of the observation period. Nevertheless, the SCOPE statistics were insignificant at the expansion point and the annual estimates had very strong upward or downward trends with some extremely high values at both ends of the period. Finally, the effort to estimate the degree of output-specific economies of scale failed in all models. A large percentage of the estimated S_k statistics had the a priori incorrect negative sign.

The estimates of cost complementarity (CC) and economies of scope (SCOPE) acquired the following values (with asymptotic standard errors in parentheses) at the expansion point:

Model	CC		SCOPE	
1. Unconstrained GTL	0.02	(0.02)	-	-
2. $\lambda_1=\lambda_2$	-0.49	(1.88)	1.75	(7.26)
3. $\lambda_T=0$	-0.17	(0.47)	0.51	(1.98)
4. $\lambda_1=\lambda_2$; $\lambda_T=0$	-0.45	(2.05)	1.27	(5.26)

None of the estimates is significantly different from zero.

Since the number of interaction terms in the 2-output TL/GTL model was still very large, multicollinearity could not be ruled out as a serious concern. In fact, the estimation problems of cost complementarity, economies of scope and especially output-specific economies of scale might be due to the still highly general nature of the specification. A comparison of the parameter estimates of the six 2-output models revealed a certain degree of instability in the first and second order output and the output-technology interaction terms. The parameters associated with these variables showed considerable fluctuation across models, some changed signs and they were generally insignificant. These variables play an important role in the calculation of test statistics for internal economies. The 2-output models seemed to suggest that some improvement with respect to the stability of parameter estimates could be achieved if the number of output-related terms were reduced.

3.3 Truncated Two-Output GTL Models

Several truncated models were estimated. The truncated models were systematically arrived at by the exclusion of parameters with insignificant t-statistics, while at the same time ensuring no significant decrease in the log of the likelihood function. Through this approach, those terms in the cost function which were not adding a significant amount of information to the overall model were excluded. The GTL model with the $\lambda_T=0$ constraint was chosen over others, partly because of its favourable economic properties in comparison with the unconstrained GTL model, and partly on the basis of likelihood ratio tests, which show in Table A.5 that three further restricted alternative models were rejected at the .05 level.

The first truncated models attempted to eliminate interaction terms associated with the toll output (Q_2), since its respective first order parameter (α_{Q_2}) was found to be consistently insignificant. The eliminated parameters were related to the toll-technology interaction term ($\beta_{Q_2}=0$), then to the squared toll variable ($\delta_{22}=0$), and finally, both parameters were excluded ($\beta_{Q_2}=\delta_{22}=0$). The insignificant output interaction term (δ_{12}) was left in the model to make the estimation of cost complementarity possible. The truncated models could not be rejected on the basis of likelihood ratio tests.

In general, the new estimates were quite stable with a few evident improvements. The Box-Cox transformation parameter for local output (λ_1) was significant in each case, as was the local-technology interaction term (β_{Q_1}) in two of the three truncated models. The first order toll output parameter (α_{Q_2}) remained insignificantly greater than zero, however, and the output interaction term (δ_{12}) and the local squared term (δ_{11}) continued to fluctuate, giving further indication of a lack of robustness with respect to these parameters.

The expansion point estimates of scale economies, cost complementarity and output-specific scale and scope economies for each model are listed below. The scale elasticity estimates exhibited the same overall pattern as in the full λ_T=0 model, although they were generally slightly higher in the truncated models. Significant cost complementarity was suggested throughout the 1970's when β_{Q_2} was set at zero. However, in the other two truncated models the cross-product derivatives were positive and significantly different from zero, indicating diseconomies for more than the second half of the observation period. The estimates of scope economies were inconsistent. Implausible results were obtained for estimates of output-specific economies of scale when β_{Q_2}=0. However, the same estimates fell into a reasonable range for both local and toll outputs in the other two models (δ_{22}=0 and β_{Q_2}=δ_{22}=0) and, as the table reveals, the local-specific economies of scale estimate was significantly greater than one at the expansion point, when δ_{22}=0. The relationship between firm level and output-specific scale elasticities, see equation (19), suggests some economies

ESTIMATES OF INTERNAL ECONOMIES IN TRUNCATED
TWO-OUTPUT GTL(λ_T=0) COST MODELS

PROPERTIES	β_{Q2} = 0	δ_{22} = 0	β_{Q2} = δ_{22} = 0
Scale	1.47* (.11)	1.45* (.11)	1.53* (.14)
Cost Complementarity	-0.58 (.34)	0.29 (.10)	0.13 (.07)
Scope	6.50 (20.4)	-0.03 (.08)	.18 (.10)
Output-Specific Scale (Local)	-8.29 (32.0)	1.50* (.24)	1.25 (.21)
Output-Specific Scale (Toll)	-5.54 (9.2)	1.31 (.21)	1.28 (.17)

Estimates are shown at the expansion point.
Asymptotic standard errors are in parentheses.
*Significant at the .05 level.

of scope - less in the $\delta_{22}=0$ model and somewhat more in the $\beta_{Q_2}=\delta_{22}=0$ model.

Further experiments restricted the output interaction parameter (δ_{12}), along with other similarly insignificantly estimated parameters, to zero. Two of these further truncated models are presented below.

The first model restricted the local squared term (δ_{11}), the output interaction term (δ_{12}) and the toll squared term (δ_{22}) to zero and the second model set the toll-technology interaction term (β_{Q_2}) to zero in addition to δ_{11}, δ_{12} and δ_{22}. These two truncated models are the end products of less restricted attempts and mark the limit to which this type of approach can be taken. Both models were narrowly rejected over the full $\lambda_T=0$ model, based upon likelihood ratio tests at the .05 level of confidence. (However, neither truncated model could be rejected at the .01 level).

In the first case, the toll parameter (α_{Q_2}) and the toll-technology interaction term (β_{Q_2}) were insignificant. After dropping the latter term in the second model, however, α_{Q_2} became significant at a value slightly greater than twice that found in the full model. Therefore, in general, it appears that reduction in the second order output-related terms results in a more efficient utilization of the available information.

Estimates of scale and scope economies at the expansion point are shown below. The scale economies once again were somewhat greater over the entire sample period than those obtained in the full $\lambda_T=0$ model. The degree of scale economies during the fifties remained low and insignificantly different from one, the estimated values in the late seventies were much higher (1.47 and 1.45 in 1978), and constant returns to scale could be rejected from 1963 onward to 1978 in each case. The estimated scope economies were quite stable over the sample period. The point estimates indicated the existence of economies of scope,

ESTIMATES OF INTERNAL ECONOMIES IN TRUNCATED
TWO-OUTPUT GTL(λ_T=0) COST MODELS

PROPERTIES	$\delta_{11} = \delta_{12} = \delta_{22} = 0$	$\delta_{11} = \delta_{12} = \delta_{22} = \beta_{Q2} = 0$
Scale	1.62* (.14)	1.62* (.13)
Scope	0.24 (.14)	0.09 (.10)
Output-Specific Scale (Local)	1.26 (.31)	1.56* (.25)
Output-Specific Scale (Toll)	1.10 (.15)	1.09 (.16)

Estimates are shown at the expansion point.
Asymptotic standard errors are in parentheses.
*Significant at the .05 level.

but they were not statistically significant. The output-specific economies of scale estimates fell into reasonable ranges in each model. However, all estimates at the expansion point were statistically insignificant. The estimates of local-specific scale economies were trended upwards in each case, reaching a maximum of 1.95 in 1978 in the second model. The degree of toll-specific scale economies exhibited very little variation, and tended to dip below one in the last years. The relationship between the respective firm level and output-specific scale economies further suggested the presence of scope economies

The two final truncated models offered many improvements over less restricted models, including the full λ_T=0 model, although the overall stability of the results was still questionable. Across the full and truncated models, wide variations in the results concerning internal economies were found under different parametric restrictions. The only consistent results observed over all the truncated models were the estimates of increasing returns to scale. To a lesser extent, especially with regards to the latter two models, a weak indication of economies of scope appeared.

3.4 Models with One Output

Another method of reducing the number of output-related terms in the TL/GTL model, alternative to truncating the function, is the reduction of the number of outputs from two to one by aggregating the local and toll service outputs of Bell Canada. The opportunity to obtain evidence on cost complementarity, economies of scope and output-specific economies of scale was lost in the resulting single output specifications, but it was hoped that greater stability of the parameters and improvement in the precision with which they would be estimated might enhance the estimates of overall economies of scale.

The unconstrained single output GTL function presented significant improvements in the estimation results. Its parameter estimates and economic properties are reported in Appendix B. The model produced a very good fit, most parameters were significant, the cost function was well behaved at most observation points and the estimated economic properties were generally realistic. A sharp estimate of scale elasticity was obtained at the expansion point. Substantial economies of scale were indicated in each year after 1956. The annual scale elasticities composed a very realistic pattern. The model strongly indicated that Bell had homothetic technology (see the likelihood ratio test in Table B.1).

Since the hypothesis of λ_Q=0 could not be rejected (see Table B.1), the output variable was introduced in logarithmic form. This led to a further improvement in the precision of the scale elasticity estimates.

The re-estimated model also proved to be homothetic. The homothetic model was tested against the homothetic $\lambda_Q \neq 0$ model and the hypothesis of $\lambda_Q=0$ could not be rejected. All parameters were significant and the estimated economic properties of the model were generally very realistic. The estimation results are described in detail in Appendix B.

The single output GTL cost function was also estimated with a logarithmic technology variable, even though the hypothesis of $\lambda_T=0$ was rejected, in order to obtain further evidence on the sensitivity of scale elasticity estimates to changes in the underlying assumptions. One model used the Box-Cox transformation of the output variable and a second model was a translog. Only slight changes and no improvements were noticed in the estimation results. As the logarithmic transformation of the technology variable generated some changes (especially towards the end of the period) in the estimates of the technology elasticity of cost, the scale elasticities became lower than in the $\lambda_T \neq 0$ models. However, their pattern did not change noticeably.

In summary, the single-output TL/GTL cost functions produced sharp and stable estimates of overall economies of scale in Bell Canada. Based on test results on various hypotheses, the homothetic GTL model in which output was logarithmically transformed appeared to be preferable to other models. Annual scale elasticity estimates from single output GTL models are shown in Appendix B. The expansion point values, together with their asymptotic standard errors are as follows:

Model	ε*
1. Unconstrained GTL	1.73 (.09)
2. $\rho_{iQ}=0$ (i=1,2,3)	1.73 (.09)
3. $\lambda_Q=0$	1.75 (.06)
4. $\lambda_Q=0$, $\rho_{iQ}=0$ (i=1,2,3)	1.73 (.05)
5. $\lambda_T=0$	1.61 (.07)
6. $\lambda_T=0$, $\rho_{iQ}=0$ (i=1,2,3)	1.69 (.09)
7. TL($\lambda_Q=\lambda_T=0$)	1.62 (.06)
8. TL($\lambda_Q=\lambda_T=0$), $\rho_{iQ}=0$ (i=1,2,3)	1.69 (.06)

*All estimates are significantly greater than one at the .05 level.

Further evidence on the robustness of scale elasticity estimates in the preferred GTL cost model (No. 4 above) was obtained by re-estimating it on slightly different data samples and with relatively minor changes in the measurement of the technology index and the cost of capital.

Sensitivity tests with respect to the technology variable were carried out first. All estimations presented above used the T2 variable in Table C.1 of Appendix C, because this measure produced generally superior empirical results. When the T3 and T4 indices of Table C.1 were substituted for T2, the scale elasticity estimate at the expansion point remained unchanged (with T3) or increased very slightly (with T4). The FNEW3 variable of Table C.1 was substituted for T2 in the next experiment. This represented a dramatic change in the measurement of the technology index, as T2 had grown almost twice as fast as FNEW3 and its pattern was also considerably different (faster growth during the first half of the period). Predictably, the scale elasticity increased significantly at the expansion point, indicating that the estimate would be sensitive to major measurement errors in the technology index. It was also indicated that major errors would destroy the reasonableness of the estimates, as the technology elasticity of cost acquired the wrong sing for the majority of the data points and other estimation problems were also encountered.

The second sensitivity test involved the capital price variable. The user cost of capital (see Kiss (1981), Appendix A, Section 3.22) was altered in three ways. First, the assumption of zero capital gain ($\dot{q}_t=0$) was relaxed. Second, the cost of common equity was altered by substituting expected yield for actual yield and changing the growth factor[14] in its formula. Third, the average cost of debt was substituted for the cost of new debt. The alternative user cost of capital measurements resulted in statistically insignificant increases in the expansion point estimate of scale elasticity.

The assumption of full capital stock utilization was relaxed in a third set of sensitivity tests. The unavailability of information and some conceptual difficulties made it impossible to approximate either the level or the annual rates of change of utilization rates. Thus, a very low (30%) and a very high (70%) level of utilization were arbitrarily chosen for the expansion point (1967). The estimated scale elasticities showed only negligible changes when either of the two utilization levels was kept constant for the entire sample period. Since improvement over time in utilization rates might be responsible for some of the estimated economies of scale, a 1% annual improvement was superimposed on the chosen expansion point levels of utilization. In order to carry the experiment to the extreme, annual improvements of 4% and 8% (for the 30% level only) were also assumed. Capital stock was rescaled by the alternative computed annual utilization rates in the total cost calculation and capital price was kept unchanged. The results can be summarized in three points. First, the scale elasticity estimates were significantly reduced but remained high (1.52 to 1.60) at the expansion point, when 1% annual improvement was assumed. Secondly, the concavity conditions were violated at all observation points and other estimation problems were encountered under higher improvement rates. Thirdly, the expansion point scale elasticity remained significantly greater than one even in the most extreme cases. It can be concluded that if Bell Canada's capital utilization has indeed improved during the period of observation the improvements appear to be responsible for a small portion of the estimated scale elasticities.

Finally, the sensitivity of scale elasticity estimates to sample variation was tested by omitting years both at the beginning and at the end of the period of observation. In order to prevent a serious loss of information, only a few data points were eliminated. When the function was re-estimated for various sub-periods, the scale elasticity estimates changed only very slightly at the expansion point.[15] The following scale elasticity estimates (with asymptotic standard errors in parentheses) were obtained in the sensitivity runs:

	MODEL		ε**
1.	Base Run:	GTL, $\lambda_Q=0$, $\rho_{iQ}=0$	1.73 (.05)
2.	Technology:	T3	1.73 (.06)
3.		T4	1.74 (.06)
4.		FNEW3	2.38 (.22)*
5.	Cost of Capital:	$\dot{q}_t \neq 0$	1.80 (.06)
6.		cost of equity	1.76 (.06)
7.		cost of debt	1.75 (.05)
8.	Cap. utilization:	30% in 1967, 0% growth	1.71 (.08)
9.		1%	1.60 (.09)*
10.		4%	1.68 (.16)
11.		8%	1.52 (.19)*
12.		70% in 1967, 0% growth	1.72 (.06)
13.		1%	1.52 (.06)*
14.		4%	1.23 (.06)*
15.	Period:	1954-1978	1.72 (.06)
16.		1956-1978	1.79 (.09)
17.		1952-1977	1.73 (.05)
18.		1952-1976	1.68 (.04)
19.		1952-1975	1.70 (.04)

*The point estimate is outside of the 95% confidence interval of the base run estimate.

**All estimates are significantly greater than one at the .05 level.

4. SUMMARY, COMPARISON, CONCLUSIONS

Twenty-three flexible translog and generalized translog cost models of Bell Canada have been examined in this paper. The most important conclusion that can be drawn from the models is that they offer a robust indication of substantial overall economies of scale in Bell Canada. A certain pattern of economies of scale estimates emerges from the comparison of models with one, two and three outputs:

MODEL	ε*
1. 3-output	1.22-1.67
2. 2-output	1.44-1.66
3. 2-output, truncated	1.45-1.62
4. 1-output	1.61-1.75

*At the expansion point (1967).

It is interesting to observe that both the expansion point estimates of scale economies and their 95% confidence intervals substantially overlap in the 2 and 3-output models. The expansion point estimates of scale economies are consistently higher in the single-output models than in the multi-output models and the overlap is narrow (1.61 to 1.67). However, due to their low standard errors, the entire 95% probability range of single-output estimates falls within the confidence interval of the multi-output estimates.

The effect of the Box-Cox generalization of variable transformation can be analyzed in the following table of estimates of economies of scale at the expansion point:

	$\lambda_k \neq 0$	$\lambda_k = 0$		$\lambda_T \neq 0$	$\lambda_T = 0$
3-output models (k=1,2,3)					
$\lambda_1 \neq \lambda_2 \neq \lambda_3$, $\lambda_T \neq 0$	1.43	1.67	$\lambda_1 \neq \lambda_2 \neq \lambda_3$	1.43	1.43
$\lambda_1 = \lambda_2 = \lambda_3$, $\lambda_T \neq 0$	1.26		$\lambda_1 = \lambda_2 = \lambda_3$	1.26	1.22
			$\lambda_k = 0$	1.67	1.67
$\lambda_1 \neq \lambda_2 \neq \lambda_3$, $\lambda_T = 0$	1.43	1.67			
$\lambda_1 = \lambda_2 = \lambda_3$, $\lambda_T = 0$	1.22				
2-output models (k=1,2)					
$\lambda_1 \neq \lambda_2$, $\lambda_T \neq 0$	1.44	1.64	$\lambda_1 \neq \lambda_2$	1.44	1.44
$\lambda_1 = \lambda_2$, $\lambda_T \neq 0$	1.50		$\lambda_1 = \lambda_2$	1.50	1.44
			$\lambda_k = 0$	1.64	1.66
$\lambda_1 \neq \lambda_2$, $\lambda_T = 0$	1.44	1.66			
$\lambda_1 = \lambda_2$, $\lambda_T = 0$	1.44				
1-output models (k=1)					
$\lambda_T \neq 0$ (homothetic)	1.73	1.73	$\lambda_Q \neq 0$	1.73	1.69
$\lambda_T = 0$	1.69	1.69	$\lambda_Q = 0$	1.73	1.69

The Box-Cox generalization of the output transformation resulted in a greater reduction of the estimated economies of scale in the 3-output models than in the 2-output models and only a negligible effect is observable in the case of single output models. At the same time, the Box-Cox generalization of the technology variable either left the scale elasticity estimates unchanged or resulted in very small changes (usually increases) in the values of ε.

In order to compare our results to those of other econometric studies, the following summary of estimates of economies of scale at the 1967 observation point or at the mean observation (which closely corresponds to 1967) in seven externally constructed TL or GTL cost functions of Bell Canada is given:

STUDY	ε
3-output models	
1. Fuss - Waverman (1978)	1.46*
2. Denny et al (1979)	1.46
3. Fuss - Waverman (1980)	.94
2-output models	
4. Smith - Corbo (1979)	1.20
5. Breslaw - Smith (1980)	1.29
1-output models	
6. Smith - Corbo (1979)	1.29
7. Denny et al (1979)	1.58

*Re-estimated in Denny et al (1979). The original estimate was 1.06 and the hypothesis of constant returns to scale could not be rejected in the original model.

With the exception of the GTL function of Fuss and Waverman (1980), all models suggested economies of scale in Bell Canada. Fuss and Waverman found that the Box-Cox transformation of the output variables reduced (from 1.43 to .94) the economies of scale estimate. This finding is consistent with our results, even the magnitude of the reduction is similar. However, our estimates were significantly higher than those of Fuss and Waverman. There are two major differences between the models which might be responsible for the difference in the level of scale elasticity estimates. First, Fuss and Waverman assumed partial profit maximization with respect to message and other toll service outputs, while cost minimization was assumed in our models. Secondly, Fuss and Waverman assumed that technological innovations improved the quality of output and translated quality improvement to quantitative increase by assuming output augmenting technological changes, while no such assumption was made in our models.

There is some overlap between our scale elasticity estimates and those from externally conducted studies in the case of the 3-output models, but our estimates are generally considerably higher in 3-output models and always higher in models with one or two outputs.

The comparison of scale elasticities in external models with one, two and three outputs is inconclusive, except for Denny et al who found lower estimates in the 3-output model than in the single output case. As mentioned above, our models yielded the same relationship.

Most external studies failed to produce a reasonable pattern for the annual economies of scale estimates. In four of the seven models that are shown above, the estimates were very strongly trended upwards. In contrast, the model by Fuss and Waverman (1980), yielded estimates in the .9 to 1.1 range, which seems to suggest that the underlying technology is linearly homogeneous. (The Fuss - Waverman estimates were also trended after 1958 in their narrow range). Two models (the single-output and 2-output translog functions of Smith and Corbo, 1979) indicated that scale elasticities gradually, but substantially, increased from 1956 to 1964 and moderately declined after 1964. This pattern appears reasonable to the extent that the new switching technologies, whose introduction began in 1956, probably increased scale elasticity and, due partly to the increasing sphere of activities of Bell Canada (e.g., intensifying regulatory activities in the 1970's) and partly to the demand slowdown of recent years, some decline in the scale elasticities during the 1970's can be reasonably expected. These characteristics as well as the remaining unrealistic features; namely, high ε values for the first years of the sample period, an inexplicable decline from 1952 to 1956 and the relatively early peak of scale elasticity in 1964; were shared by our mutliple output models and were eliminated only in the single output models. The preferred single output homothetic GTL function with logarithmic output shows relatively stable but low values for 1952-55, replaces the 1964 peak with a flat maximum in the 1968 to 1971 period and makes the decline of scale elasticities during the 1970's less pronounced.

The second conclusion that can be drawn from the study is that statistically insignificant global and local economies of scope were generally indicated by the estimated multi-output cost models. The standard errors of estimates (as well as the annual estimates themselves) improved when the number of outputs was reduced from three to two and this suggests that the insignificance of the estimates is due to a high degree of multicollinearity in the models.

An almost perfectly uniform indication of global economies of scope was obtained for all output categories and all years in all models, where the form of the function allowed for the computation of the SCOPE_k statistic. Compared to the two well-behaved 3-output cost functions, the 2-output models reduced the standard error of SCOPE_k rather drastically and resulted in generally reasonable values for the annual estimates.

Local economies of scope were indicated by the negative sign of estimates of the cost complementarity statistic between local and toll service outputs. However, none of the computed statistics were significantly different from zero at the expansion point and the annual estimates were generally trended. All constrained 2-output models indicated complementarity (the full model produced a very small positive value) between local and total toll and the two well-behaved 3-output models gave a further suggestion that complementarity existed between local and both message and other toll. It is interesting to observe that the indication of local/other toll complementarity was very strong at the expansion point and the annual estimates were not trended.

Only one external study offers estimates of global economies of scope. Fuss and Waverman (1980) found that the annual estimates of economies of scope, specific to other toll services were downward trended and switched signs (from positive to negative) in 1962. The estimates were small and all but one were insignificant. Our comparable other toll-specific economies of scope estimates were positive for the entire period of observation, but the values were trended and fell into the reasonable range only at the beginning of the period.

The following table sums up the evidence that exists in externally conducted econometric studies of Bell Canada. The computed CC values are shown either at the expansion point of the model or in 1967.

	Fuss - Waverman (1978)	Denny et al (1979)	Fuss - Waverman (1980)	Smith - Corbo (1979)	Breslaw - Smith (1980)
Local - Message Toll	-.016	-.062*	.099	-.001	.000
Local - Other Toll	.009*	-.037*	.042	-	-
Message - Other Toll	-.002	.017*	-.021*	-	-
Local - Total Toll	-	-	-	-	-

*Significantly different from zero.

The evidence is inconclusive. The indication from our 3-output models is similar to Denny's results, but there is no further resemblance between our estimates and those shown in the table.

Our models failed to produce both reasonable and statistically significant estimates of output-specific economies of scale with any degree of consistency. In general, the estimates either had the a priori incorrect negative sign, or the positive estimates were unrealistic in magnitude and in pattern in the non-truncated models.[16] On the other hand, encouraging results were obtained from the truncated 2-output models. Two models yielded reasonable and significant estimates of local-specific economies of scale, and fully consistent estimates of overall and output-specific economies of scale and global economies of scope were obtained from three truncated models. These models suggested a higher degree of economies of scale in local than in toll services and also suggested that the degree of overall economies of scale was greater than that of either of the two output-specific economies of scale; thus, global economies of scope also existed in Bell Canada. Based on our estimation results, it appears that the truncation of cost functions (justified by both t-tests and likelihood ratio tests) is a promising method of breaking the multicollinearity of multi-output models and obtaining econometric evidence on economies of scope and output-specific economies of scale.

To summarize our conclusions, a robust indication of substantial overall economies of scale was obtained from estimated cost models of Bell Canada. The evidence these models offer on economies of scope is uniform but not as convincing, either from a statistical or from an economic point of view. Finally, the difficulties associated with the estimation of multiple output cost models and the possible hazards of estimating hypothetical production cost for zero output levels prevented us from producing strong econometric evidence on economies of scale with respect to Bell Canada's local and toll services. However, the statistically justified truncation of cost functions yielded encouraging results.

FOOTNOTES

[1] *Note that the term "evidence" has no legal or regulatory reference. It denotes the body of empirical knowledge, obtained from econometric studies.*

[2] *One advantage is that output is an exogenous variable in the cost function. This makes cost functions especially suitable for regulated public utilities, whose prices are determined by regulatory agencies, hence their output levels are exogenously set. Another advantage is the relative ease with which the multi-output case can be considered in cost models. Multi-output specifications are required for obtaining evidence on cost complementarity, economies of scope and output-specific economies of scale. Still another advantage of cost functions is that they yield direct estimates of such important properties as marginal costs, cost elasticities and partial elasticities of factor substitution. The most notable disadvantages are the problems of measuring the cost of capital and the lack of direct marginal product estimates.*

[3] *The translog cost function was first used to estimate economies of scale by Christensen and Greene (1976) and was used in empirical studies of Bell Canada by Fuss and Waverman (1977, 1978), Denny et al (1979), Smith and Corbo (1979), Breslaw and Smith (1980) and also in studies of the Bell System by Nadiri and Shankerman (1979) and Christensen et al (1980). The Box-Cox transformation (Box and Cox, 1962) of variables of cost functions was proposed by Khaled (1978) and the first GTL of Bell Canada was estimated by Fuss and Waverman (1980).*

[4] *Total cost and input prices remain in logarithmic form in GTL in order to facilitate easy parameter restrictions to ensure the first degree homogeneity of the cost function in input prices. See the restrictions in (3).*

[5] *T* is introduced as a variable in the expansion of the production frontier. The specification allows for input-neutral as well as biased changes in technology. The same solution can be seen in a number of sources, e.g., Smith and Corbo (1979). In contrast, Fuss and Waverman (1977, 1978, 1980) and Denny et al (1979) assumed input or output augmenting technological changes in their multiple output models.*

[6] *The restriction to first degree homogeneity ensures that when all input prices are changed by the same percentage and outputs and technology remain unchanged, the resulting percentage change in total cost will be equal to the equiproportional input price change.*

[8] *There are several reasons for regarding exogenous toll output as a more realistic assumption than endogenous toll output for Bell Canada. First, the overwhelming majority of toll rates are set basically exogenously; i.e., influenced to various degrees, but not determined by Bell Canada. This applies to all message toll service categories. Intra-Bell rates are regulated, while TCTS, adjacent member, US and overseas long distance tariffs are set, usually for several years, by bilateral and multilateral agreements and are subject to approval by the CRTC. Secondly, the assumption that Bell Canada reaps maximum monopoly profit on services whose prices are endogeneously set appears to be unrealistic, because of of the competitive nature of these services. Thirdly, the internal procedures of Bell Canada reflect the company's ambition to minimize production costs but do not reveal any pursuit of a monopolistic profit maximum. The company's budgeting process and the procedures aimed at determining the level of rate increases do not contain calculations equating the marginal costs and marginal revenues of services. Finally, there are indications (see Bell Canada General Increase in Rates (1980), Exhibit Nos. B-80-200 (Section 5) and B-80-234) that demand at least for the largest categories of toll services may be price inelastic. Monopolistic profit maximum cannot exist in the region of price inelastic demand.*

[9] *For more on the problems of measuring economies of scale, and some conceptual clarification, see Hanoch (1975).*

[10] *The restrictions in (9) imply constant returns to scale at each observation point in TL, but the GTL function is restricted by them to exhibit constant returns to scale at the expansion point only, where* $Q_k=1$.

[11] *See Baumol and Braunstein (1977).*

[12] *The estimates of output-specific economies of scale and scope were either negative or had very large positive values. However, the two well-behaved cost functions indicated cost complementarity between local and other toll services for each year of the observation period and the annual estimates were not trended. An inconsistent indication of cost complementarity was obtained with respect to local and message toll services, while no complementarity was evident between message and other toll services.*

[13] *The hypothesis of homotheticity could not be rejected in the* (λ_k=0, k=1,2) *model. The estimated scale elasticity remained virtually unchanged at the expansion point and underwent only very small changes at the two end points of the sample in the homothetic model. The cost complementarity statistic retained its negative sign. However, the homothetic function produced some unreasonable (negative) estimates of the marginal cost of toll service output.*

[14] *Expected yield is the annualized quarterly dividend in percent of the average market price of common shares, declared in the fourth quarter of the previous year. Actual yield is the declared dividend relative to the actual average market price of common shares in the test year. The 10-year loglinear average growth rate of dividends and earnings per share was substituted for that of dividends alone.*

[15] *Small changes were registered in the scale elasticity estimates at the begining of the observation period. The 1975 estimate of* ε *increased gradually as the last year, the last two years and the last three years were omitted in successive steps.*

[16] *In the 3-output GTL cost model of Fuss and Waverman (1980), the hypothesis of constant returns to scale with respect to other toll services was rejected and the annual estimates of economies of scale of other toll fell in the 2.12 to 2.37 range.*

REFERENCES

Baumol, W.J. and Y.M. Braunstein, "Empirical Study of Scale Economies and Production Complementarity: The Case of Journal Publication" Journal of Political Economy, October 1977, pp. 1037-48.

Baumol, W.J., D. Fischer and M.I. Nadiri, "Forms of Empirical Cost Functions to Evaluate Efficiency of Industry Structure", Paper No. 30, Centre for the study of Business Regulation, Graduate School of Business Administration, Duke University, 1978.

Bell Canada General Increase in Rates, 1980, Part B: Memoranda of Support, February 19, 1980.

Box, G.E.P. and D.R. Cox, "An Analysis of Transformations (with Discussions)", Journal of the Royal Statistical Society, Series B, 1962, pp. 211-243.

Breslaw, J. and B. Smith, "Efficiency, Equity and Regulation: An Econometric Model of Bell Canada", Final Report to the Department of Communications, March 1980.

Christensen, L.R., D. Cummings and P.E. Schoech, "Productivity in the Bell System 1947-1977", paper presented at the Eighth Annual Telecommunications Policy Research Conference, April 27-30, 1980.

Christensen, L.R. and W.H. Greene, "Economies of Scale in U.S. Electric Power Generation", Journal of Political Economy, August 1976, pp. 655-676.

Corbo, V., J. Breslaw, J.M. Dufour, and J.M. Vrljicak, "A Simulation Model for Bell Canada: Phase II", Institute of Applied Economic Research Concordia University, March 1979.

Denny, M., M. Fuss, and C. Everson, "Productivity, Employment and Technical Change in Canadian Telecommunications: The Case for Bell Canada", Final Report to the Department of Communications, March 1979.

Fuss, M. and L. Waverman, "Multi-Product, Multi-Input Cost Functions for a Regulated Utility: The Case of Telecommunications in Canada", paper presented at the N.B.E.R. Conference on Public Regulation, Washington December 1977.

____________________, "Multi-Product Multi-Input Cost Functions for a Regulated Utility: The Case of Telecommunication in Canada", Draft, Institute for Policy Analysis, June 1978, revision of Fuss and Waverman (1977).

____________________, "The Regulation of Telecommunications in Canada", Draft Copy of the Final Report to the Economic Council of Canada, June 25, 1980.

Hanoch, G. "The Elasticity of Scale and the Shape of the Average Costs", The American Economic Review, June 1975, pp. 492-496.

Khaled, M.S. "Productivity Analysis and Functional Specification: A Parametric Approach". Ph.D. dissertation, University of British Columbia, Department of Economics, April 1978.

Kiss, F. "The Bell Canada Productivity Study", Corporate Analysis Division, Bell Canada, February 1981.

Nadiri, M.I. and M.A. Schankerman, "The Structure of Production, Technological Change and the Rate of Growth of Total Factor Productivity in the Bell System", New York University, National Bureau of Economic Research, 1979.

Panzar, J.C. and R.D. Willig, "Economies of Scope, Product Specific Economies of Scale, and the Multi-Product Competitive Firm", Bell Laboratories, 1978. Revised in 1979.

Shephard, R.W., Cost and Production Functions. Princeton University Press, Princeton, N.J., 1970.

Smith, J.B., and V. Corbo, "Economies of Scale and Economies of Scope of Bell Canada", Institute of Applied Economic Research, Concordia University, Final Report to the Department of Communications, March 1979.

Willig, R.D. "Multi-Product Technology and Market Structure", The American Economic Review, Vol. 69, No. 2, May 1979, pp. 346-351.

Zellner, A., "An Efficient Method for Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias", Journal of the American Statistical Association, vol. 57, June 1962, pp. 348-368.

APPENDIX A

ESTIMATES OF TWO-OUTPUT TL AND GTL COST MODELS

TABLE A.1: R^2 AND DURBIN-WATSON STATISTICS IN TWO-OUTPUT TL AND GTL COST MODELS

MODEL	COST FUNCTION R^2	COST FUNCTION DW	LABOUR SHARE R^2	LABOUR SHARE DW	CAPITAL SHARE R^2	CAPITAL SHARE DW
GTL; unconstrained	.9993	1.49	.9943	1.15	.9936	1.39
$\lambda_Q=\lambda_{Q1}=\lambda_{Q2}$	.9991	1.37	.9957	1.86	.9931	1.41
$\lambda_Q=0$	.9992	1.33	.9918	1.18	.9912	1.30
$\lambda_T=0$	.9992	1.33	.9961	1.58	.9932	1.31
$\lambda_Q=\lambda_{Q1}=\lambda_{Q2}$; $\lambda_T=0$	.9978	1.36	.9646	1.84	.9595	1.40
TL; $\lambda_Q=\lambda_T=0$	.9978	1.34	.9499	1.27	.9534	1.35

TABLE A.2: CORRELATION COEFFICIENTS OF ACTUAL TO FITTED FACTOR INPUTS IN TWO-OUTPUT TL AND GTL COST MODELS

MODEL	LABOUR	CAPITAL	MATERIAL
GTL; unconstrained	.9917	.9997	.9953
$\lambda_Q=\lambda_{Q1}=\lambda_{Q2}$	.9936	.9996	.9954
$\lambda_Q=0$	.9871	.9994	.9956
$\lambda_T=0$	.9945	.9996	.9948
$\lambda_Q=\lambda_{Q1}=\lambda_{Q2}$; $\lambda_T=0$	.9939	.9996	.9953
TL; $\lambda_Q=\lambda_T=0$	.9873	.9994	.9956

TABLE A.3: PARAMETER ESTIMATES OF TWO-OUTPUT TL AND GTL COST MODELS (ASYMPTOTIC STANDARD ERRORS IN PARENTHESES)

PARA- METERS	GTL unconstrained		GTL $\lambda_0=\lambda_1=\lambda_2$		GTL $\lambda_k=0$	
α_0	.015	(.004)	.017	(.005)	.010	(.005)
α_1	.318	(.002)	.318	(.002)	.318	(.002)
α_2	.510	(.002)	.511	(.002)	.511	(.002)
α_3	.171	(.002)	.171	(.002)	.171	(.002)
α_{Q1}	.633	(.039)	.565	(.151)	.502	(.145)
α_{Q2}	.063	(.036)*	.102	(.089)*	.109	(.098)*
β	-.204	(.082)	-.280	(.083)	-.124	(.077)*
γ_{11}	.063	(.019)	.092	(.026)	.123	(.030)
γ_{12}	-.109	(.010)	-.142	(.011)	-.160	(.014)
γ_{13}	.046	(.016)	.050	(.023)	.038	(.024)*
ρ_{11}	.009	(.007)*	-.257	(.056)	-.129	(.080)*
ρ_{12}	.002	(.002)*	.157	(.032)	.083	(.052)*
β_1	-.178	(.015)	-.192	(.028)	-.213	(.033)
γ_{22}	.214	(.007)	.236	(.009)	.238	(.011)
γ_{23}	-.105	(.011)	-.093	(.012)	-.078	(.014)
ρ_{21}	-.002	(.008)*	.192	(.064)	.071	(.086)*
ρ_{22}	-.002	(.001)*	-.118	(.040)	-.045	(.059)*
β_2	.216	(.016)	.239	(.033)	.243	(.039)
γ_{33}	.059	(.018)	.044	(.026)*	.040	(.025)*
ρ_{31}	.011	(.007)*	.065	(.052)*	.058	(.041)*
ρ_{32}	-.001	(.001)*	-.039	(.026)*	-.038	(.024)*
β_3	-.038	(.015)	-.047	(.021)	-.029	(.016)*
δ_{11}	.306	(.103)	.687	(3.614)*	3.782	(2.467)*
δ_{12}	-.016	(.031)*	-.551	(1.885)*	-1.908	(1.376)*
δ_{22}	.003	(.004)*	.526	(1.831)*	1.076	(.768)*
β_{Q1}	-.329	(.212)*	-1.821	(.987)*	-1.832	(.886)
β_{Q2}	-.328	(.168)	-.228	(1.018)*	.374	(.585)*
β_T	1.127	(.318)*	3.159	(.803)	2.098	(.700)
λ_T	-.532	(.154)	.283	(.375)*	-.238	(.435)*
λ_Q			.602	(.123)		
λ_1	-1.053	(.168)				
λ_2	3.214	(.805)				

TABLE A.3 (Cont'd)

PARAMETERS	GTL $(\lambda_T=0)$		GTL $(\lambda_Q=\lambda_1=\lambda_2;\ \lambda_T=0)$,		TL $(\lambda_k=\lambda_T=0)$	
α_0	.015	(.005)	.017	(.005)	.011	(.004)
α_1	.318	(.002)	.319	(.002)	.319	(.002)
α_2	.511	(.002)	.511	(.002)	.511	(.002)
α_3	.170	(.002)	.171	(.002)	.171	(.002)
α_{Q1}	.650	(.095)	.624	(.152)	.463	(.124)
α_{Q2}	.046	(.053)*	.069	(.087)*	.141	(.080)*
β	-.283	(.082)	-.288	(.082)	-.151	(.070)
γ_{11}	.061	(.025)	.087	(.026)	.128	(.028)
γ_{12}	-.136	(.011)	-.143	(.011)	-.162	(.013)
γ_{13}	.075	(.023)	.056	(.023)	.034	(.023)*
ρ_{11}	-.122	(.052)	-.228	(.047)	-.166	(.040)
ρ_{12}	.071	(.028)	.137	(.022)	.109	(.023)
β_1	-.167	(.016)	-.177	(.015)	-.230	(.017)
γ_{22}	.233	(.009)	.237	(.009)	.236	(.011)
γ_{23}	-.097	(.012)	-.095	(.012)	-.075	(.013)
ρ_{21}	.087	(.044)	.161	(.053)	.111	(.043)
ρ_{22}	-.047	(.023)	-.095	(.028)	-.074	(.025)
β_2	.218	(.023)	.222	(.021)	.263	(.018)
γ_{33}	.022	(.025)*	.038	(.026)*	.041	(.025)*
ρ_{31}	.035	(.038)*	.067	(.057)*	.055	(.038)*
ρ_{32}	-.024	(.016)*	-.041	(.028)*	-.035	(.022)*
β_3	-.051	(.023)	-.045	(.020)	-.033	(.016)
δ_{11}	.219	(1.10)*	.652	(3.93)*	3.185	(2.00)*
δ_{12}	-.204	(.480)*	-.495	(2.06)*	-1.617	(1.15)*
δ_{22}	.209	(.220)*	.434	(1.09)*	.971	(.683)*
β_{Q1}	-.850	(.962)*	-2.08	(1.64)*	-1.639	(.851)*
β_{Q2}	-.358	(.394)*	.002	(.837)*	.160	(.463)*
β_T	1.763	(.710)	2.783	(.694)	2.391	(.523)
λ_Q			.672	(.087)		
λ_1	.342	(.251)*				
λ_2	.754	(.132)				

* Not significant at the .05 level.

TABLE A.4: SCALE ELASTICITY ESTIMATES IN TWO-OUTPUT TL AND GTL COST MODELS

YEAR	GTL Unconstrained	GTL $\lambda_Q=\lambda_1=\lambda_2$	GTL $\lambda_k=0$	GTL $\lambda_T=0$	GTL $\lambda_Q=\lambda_1=\lambda_2$ $\lambda_T=0$	TL $\lambda_k=\lambda_T=0$
1952	1.21	1.04	1.12	1.05	1.00	1.14
1953	1.02	1.00	1.06	1.03	.97	1.09
1954	.92	.98	1.01	1.01	.93	1.04
1955	.84	.92	.97	.98	.88	1.00
1956	.84	.90	.97	.96	.86	.99
1957	.86	.88	.95	.96	.84	.97
1958	.93	.94	1.01	1.02	.90	1.04
1959	.97	.96	1.06	1.05	.92	1.08
1960	1.05	1.04	1.14	1.12	1.00	1.17
1961	1.10	1.09	1.16	1.16	1.05	1.19
1962	1.14	1.11	1.25	1.18	1.07	1.27
1963	1.23	1.28	1.43	1.30	1.24	1.46
1964	1.32	1.53	1.78	1.44	1.48	1.79
1965	1.36	1.54	1.77	1.45	1.49	1.78
1966	1.41	1.54	1.72	1.45	1.48	1.73
1967	1.44	1.50	1.64	1.44	1.44	1.66
1968	1.51	1.62	1.70	1.50	1.56	1.72
1969	1.55	1.67	1.74	1.52	1.60	1.76
1970	1.61	1.71	1.66	1.54	1.63	1.70
1971	1.68	1.75	1.52	1.55	1.66	1.59
1972	1.69	1.73	1.46	1.53	1.64	1.53
1973	1.59	1.63	1.50	1.47	1.57	1.55
1974	1.51	1.58	1.44	1.42	1.53	1.49
1975	1.31	1.53	1.40	1.35	1.48	1.45
1976	1.11	1.49	1.32	1.30	1.45	1.39
1977	.87	1.42	1.28	1.23	1.39	1.35
1978	.63	1.42	1.31	1.18	1.40	1.38

TABLE A.5: LIKELIHOOD RATIO TEST RESULTS IN TWO-OUTPUT GTL COST MODEL (NO. OF RESTRICTIONS IN PARENTHESES)

UNCONSTRAINED MODELS CONSTRAINED MODELS	GTL; Unconstrained	$\lambda_1=\lambda_2$	$\lambda_T=0$	$\lambda_1=\lambda_2=0$	$\lambda_1=\lambda_2$ $\lambda_T=0$
$\lambda_1=\lambda_2$	14.33 (1)	-			
$\lambda_T=0$	8.27 (1)	-	-		
$\lambda_1=\lambda_2=0$	35.32 (2)	21.00 (1)	27.05 (2)	-	
$\lambda_1=\lambda_2$; $\lambda_T=0$	15.10 (2)	0.77 (1)	6.06 (1)	-	-
TL; $\lambda_1=\lambda_2=\lambda_T=0$	35.70 (3)	21.37 (2)	27.43 (2)	0.38 (1)	20.60 (1)

NOTE 1: The $\chi^2_{.05,r}$ ($\chi^2_{.01,r}$) critical values for r = 1,2,3 are 3.84 (6.63), 5.99 (9.21) and 7.81 (11.34) respectively.

NOTE 2: The test results in column 1 indicate that the unconstrained GTL model cannot be rejected in favour of any of the five constrained models. However, this is only weakly suggested in the case of the second null hypothesis, $\lambda_T=0$.

over

NOTE 3: If $\lambda_T=0$ is assumed, contrary to NOTE 2 but justified to some extent by the better economic properties of the so restricted model, the resulting model cannot be rejected against any of the further constrained models at the .05 level of confidence. However, the null hypothesis that $\lambda_1=\lambda_2$ is narrowly accepted at the 0.1 level. See column 3.

NOTE 4: When the $\lambda_1=\lambda_2$ assumption is made, even if it does not necessarily follow from NOTE 3, the test indicates the acceptance of the $\lambda_T=0$ hypothesis and the result is the TL specification. See column 4.

TABLE A.6: LIKELIHOOD RATIO TEST RESULTS IN TRUNCATED TWO-OUTPUT GTL ($\lambda_T=0$) MODELS

RESTRICTION	NO. OF RESTRICTIONS	$X^2_{.05,r}$	$X^2_{.01,r}$	TEST VALUE
$\beta_{Q_2}=0$	1	3.84	-	0.59
$\delta_{22}=0$	1	3.84	-	0.51
$\beta_{Q_2}=\delta_{22}=0$	2	5.99	-	2.73
$\delta_{11}=\delta_{12}=\delta_{22}=0$	3	7.81	11.34	10.25
$\delta_{11}=\delta_{12}=\delta_{22}=\beta_{Q_2}=0$	4	9.49	13.28	11.55

TABLE A.7: PARAMETER ESTIMATES OF TRUNCATED TWO-OUTPUT GTL (λ_T=0) MODELS (ASYMPTOTIC STANDARD ERRORS IN PARENTHESES)

PARA- METERS	$\beta_{Q2} = 0$		$\delta_{22} = 0$		$\beta_{Q2} = \delta_{22} = 0$	
α_0	.015	(.005)	.015	(.005)	.013	(.005)
α_1	.319	(.002)	.319	(.002)	.319	(.002)
α_2	.511	(.002)	.511	(.002)	.511	(.002)
α_3	.170	(.002)	.170	(.002)	.170	(.002)
α_{Q1}	.626	(.098)	.651	(.102)	.610	(.111)
α_{Q2}	.057	(.054)*	.038	(.056)*	.042	(.061)*
β	-.259	(.083)	-.260	(.082)	-.170	(.080)
γ_{11}	.064	(.025)	.062	(.025)	.062	(.026)
γ_{12}	-.137	(.011)	-.140	(.010)	-.144	(.011)
γ_{13}	.073	(.021)	.078	(.022)	.082	(.020)
ρ_{11}	-.138	(.046)	-.138	(.051)	-.156	(.046)
ρ_{12}	.080	(.025)	.080	(.027)	.093	(.025)
β_1	-.167	(.016)	-.167	(.016)	-.166	(.016)
γ_{22}	.232	(.009)	.237	(.009)	.237	(.010)
γ_{23}	-.095	(.012)	-.097	(.012)	-.093	(.012)
ρ_{21}	.099	(.044)	.100	(.045)	.121	(.048)
ρ_{22}	-.054	(.022)	-.054	(.024)	-.066	(.024)
β_2	.217	(.023)	.216	(.023)	.215	(.023)
γ_{33}	.022	(.024)*	.019	(.024)*	.011	(.023)*
ρ_{31}	.039	(.040)*	.036	(.041)*	.035	(.046)*
ρ_{32}	-.026	(.017)*	-.026	(.017)*	-.027	(.020)*
β_3	-.051	(.023)	-.050	(.023)	-.049	(.024)
δ_{11}	1.146	(.646)*	-.804	(.389)	-.174	(.294)*
δ_{12}	-.613	(.343)*	.266	(.103)	.107	(.067)*
δ_{22}	.385	(.213)*	.0		.0	
β_{Q1}	-1.657	(.226)	-.434	(.801)*	-1.659	(.226)
β_{Q2}	.0		-.567	(.329)*	.0	
β_T	2.300	(.294)	1.593	(.664)	2.461	(.301)
λ_1	.415	(.159)	.408	(.206)	.537	(.100)
λ_2	.742	(.121)	.741	(.119)	.758	(.104)

* Not significant at the .05 level.

TABLE A.7 (Cont'd)

PARA-METERS	$\delta_{11} = \delta_{12} = \delta_{22} = 0$		$\delta_{11} = \delta_{12} = \delta_{22} = \beta_{Q2} = 0$	
α_0	.011	(.005)	.011	(.005)
α_1	.319	(.002)	.319	(.002)
α_2	.512	(.002)	.512	(.002)
α_3	.169	(.002)	.169	(.002)
α_{Q1}	.521	(.099)	.503	(.085)
α_{Q2}	.096	(.051)*	.114	(.043)
β	-.151	(.077)	-.175	(.076)
γ_{11}	.046	(.025)	.042	(.023)*
γ_{12}	-.147	(.011)	-.147	(.011)
γ_{13}	.101	(.021)	.105	(.019)
ρ_{11}	-.119	(.048)	-.098	(.030)
ρ_{12}	.073	(.026)	.061	(.015)
β_1	-.165	(.016)	-.164	(.016)
γ_{22}	.222	(.007)	.220	(.008)
γ_{23}	-.075	(.012)	-.073	(.012)
ρ_{21}	.103	(.048)	.082	(.035)
ρ_{22}	-.051	(.025)	-.039	(.015)
β_2	.217	(.025)	.220	(.026)
γ_{33}	-.025	(.023)*	-.032	(.022)*
ρ_{31}	.016	(.043)*	.016	(.036)*
ρ_{32}	-.022	(.018)*	-.022	(.014)*
β_3	-.052	(.026)	-.056	(.027)
δ_{11}	.0		.0	
δ_{12}	.0		.0	
δ_{22}	.0		.0	
β_{Q1}	-1.638	(.671)	-1.136	(.194)
β_{Q2}	.212	(.270)*	.0	
β_T	2.059	(.471)	1.669	(.199)
λ_1	.488	(.190)	.326	(.098)
λ_2	.858	(.130)	.863	(.147)

* Not significant at the .05 level.

TABLE A.8: SCALE ELASTICITY ESTIMATES IN TRUNCATED TWO-OUTPUT GTL (λ_T=0) MODELS

YEARS	$\beta_{Q2}=0$	$\delta_{22}=0$	$\beta_{Q2}=\delta_{22}=0$	$\delta_{11}=\delta_{12}=\delta_{22}=0$	$\delta_{11}=\delta_{12}=\delta_{22}=\beta_{Q2}=0$
1952	1.07	1.00	.99	1.03	1.04
1953	1.03	1.00	.97	1.01	1.03
1954	1.00	.99	.95	.99	1.03
1955	.95	.97	.91	.95	1.00
1956	.93	.97	.90	.95	1.00
1957	.92	.97	.90	.95	1.01
1958	.98	1.03	.96	1.02	1.08
1959	1.01	1.06	1.00	1.06	1.12
1960	1.09	1.13	1.09	1.16	1.22
1961	1.13	1.19	1.15	1.22	1.28
1962	1.16	1.20	1.18	1.25	1.31
1963	1.31	1.31	1.35	1.43	1.47
1964	1.53	1.42	1.57	1.65	1.65
1965	1.53	1.44	1.58	1.66	1.66
1966	1.51	1.45	1.57	1.65	1.65
1967	1.47	1.45	1.53	1.62	1.62
1968	1.55	1.51	1.62	1.71	1.69
1969	1.58	1.52	1.63	1.73	1.70
1970	1.58	1.55	1.65	1.74	1.71
1971	1.56	1.61	1.67	1.76	1.72
1972	1.52	1.59	1.62	1.73	1.69
1973	1.49	1.47	1.51	1.65	1.62
1974	1.45	1.42	1.44	1.61	1.58
1975	1.40	1.34	1.36	1.56	1.53
1976	1.34	1.29	1.30	1.51	1.49
1977	1.28	1.22	1.23	1.47	1.45
1978	1.26	1.15	1.20	1.47	1.45

APPENDIX B

ESTIMATES OF SINGLE-OUTPUT GTL COST MODELS

The unconstrained GTL model was estimated first. The parameter estimates of the input-output interaction terms (ρ_{1Q}, ρ_{2Q}, ρ_{3Q}) and the Box-Cox transformation of the output term (λ_Q) were insignificantly different from zero. This suggested that the specification could be reduced to a homothetic model ($\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0$) with logarithmic transformation of output ($\lambda_Q=0$). Consequently, two independently constrained models were estimated. The first model eliminated the input-output interaction parameters ($\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0$), while in the second, the transformation of the Q variable was restricted to be logarithmic. Since the λ_Q parameter in the first restricted model and the input-output interaction terms in the second restricted model were insignificantly different from zero, the next step was to apply the hypothesis $\lambda_Q=0$; $\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0$. To ensure that no significant amount of information was lost by the imposed restrictions, likelihood ratio tests were applied at every step. Test results are presented in Table B.1. As we moved from one set of restrictions to the other, the change in the log of likelihood function proved to be marginal. In order to test whether the technology variable should be reduced to TL, the above mentioned models were estimated under the hypotheses $\lambda_T=0$ and $\lambda_Q=\lambda_T=0$. The loss in the log of likelihood function in these estimations, however, proved to be significant; thus, the models were rejected at the .05 level.

The parameter estimates of the first four models, presented in Table B.2, were found to be very stable. The behaviour characteristics, the estimated properties as well as the scale elasticities (Table B.3) followed a similar pattern. On the basis of likelihood ratio tests, the homothetic GTL with natural logarithm for the output and Box-Cox transformation for the technology variable was chosen as the reduced cost model.

As can be seen in Table B.2, all parameter estimates with the exception of the second order output parameter (δ_{QQ}) and the material-technology interaction term (β_3) are statistically significant at the .05 level. The model indicates that technological change is capital using ($\beta_2>0$), labour saving ($\beta_1<0$) and material neutral.

The cost function is well behaved. Monotonicity is satisfied at all observation points and the concavity conditions are met for 78% of the observation points.

Cost elasticities with respect to technology, shown in Table B.4, have the expected sign in each year. The absolute values are low for the first two years of the sample, peak in 1955-56 and fall until 1964, after which they are trended upward.

Table B.5 presents the own price elasticities of factor demand. Labour and material price elasticities have the correct negative sign throughout the sample period, but the capital price elasticities are positive for the first 5 years. Demand for all three productive factors is price inelastic, with material demand being the relatively most sensitive (-.5) and capital demand the most insensitive (almost perfectly inelastic at -.05) to price changes. The price elasticity estimates are stable through time.

Partial elasticities of factor substitution in Table B.6 show a high degree of labour-material substitutability and, more importantly, a low degree of substitutability between capital and labour. Both are fairly stable throughout the observation period. Capital and material are complementary in the early and mid 1950's and neither complementarity nor substitutability is indicated during the 1960's and 1970's.

TABLE B.1: LIKELIHOOD RATIO TEST RESULTS OF SINGLE-OUTPUT COST (NO. OF RESTRICTIONS IN PARENTHESES)

UNCONSTRAINED MODELS / CONSTRAINED MODELS	UNCONSTRAINED GTL	HOMOTHETIC ($\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0$)	$\lambda_Q=0$	HOMOTHETIC ($\lambda_Q=0;\rho_{1Q}=\rho_{2Q}=\rho_{Q3}=0$)
$\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0$	.58 (3)	-	n.a.	n.a.
$\lambda_Q=0$	.16 (1)	n.a.	-	n.a.
$\lambda_T=0$	10.0 (1)	n.a.	n.a.	n.a.
$\lambda_Q=0;$ $\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0$	.63 (4)	.01 (1)	.43 (3)	-
$\lambda_T=0;$ $\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0$	14.58 (4)	14.0 (1)	n.a.	n.a.
$\lambda_Q=\lambda_T=0$	10.70 (2)	n.a.	10.54 (1)	n.a.
$\lambda_Q=\lambda_T=0;$ $\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0$	14.64 (5)	14.05 (2)	14.47 (4)	14.05 (1)

NOTE 1: The $\chi^2_{.05,r}$ ($\chi^2_{.01,r}$) critical values of r=1,2,3,4,5 are 3.84(6.63), 5.99(9.21), 7.81(11.34), 9.49(13.28) and 11.07(15.09)

over

NOTE 2: The test results in column 1 suggest that the unconstrained GTL model can be rejected in favour of the homothetic GTL (λ_Q=0;ρ_{1Q}=ρ_{2Q}=ρ_{3Q}=0) and homothetic TL (λ_Q=λ_T=0;ρ_{1Q}=ρ_{2Q}=ρ_{3Q}=0); the latter at the .01 level only. However, when the assumption is the homothetic GTL (λ_Q=0; ρ_{1Q}=ρ_{2Q}=ρ_{3Q}=0), the homothetic TL is rejected at every significance level.

TABLE B.2: PARAMETER ESTIMATES OF SINGLE-OUTPUT GTL COST MODELS

(ASYMPTOTIC STANDARD ERRORS IN PARENTHESES)

PARAMETER	UNCONSTRAINED		HOMOTHETIC ($\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0$)		$\lambda_Q=0$		HOMOTHETIC ($\lambda_Q=0;\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0$)	
α_0	.014	(.005)	.013	(.005)	.013	(.004)	.013	(.004)
α_1	.318	(.002)	.317	(.002)	.318	(.002)	.317	(.002)
α_2	.510	(.002)	.570	(.002)	.510	(.002)	.510	(.002)
α_3	.171	(.002)	.172	(.002)	.171	(.002)	.172	(.002)
α_Q	.579	(.032)	.577	(.030)	.573	(.021)	.577	(.017)
λ_Q	.081*	(.248)*	-.012	(.354)*	.0		.0	
γ_{11}	.082	(.030)	.108	(.014)	.091	(.030)	.108	(.014)
γ_{12}	-.137	(.014)	-.138	(.010)	-.139	(.013)	-.138	(.010)
γ_{13}	.054	(.024)	.029	(.010)	.048	(.023)	.029	(.010)
γ_{22}	.224	(.012)	.226	(.011)	.225	(.011)	.226	(.011)
γ_{23}	-.087	(.015)	-.088	(.013)	-.086	(.014)	-.088	(.013)
γ_{33}	.033	(.027)*	.059	(.015)	.038	(.026)*	.059	(.015)
δ_{QQ}	.067	(.117)*	.100	(.189)*	.100	(.058)*	.097	(.056)*
ρ_{1Q}	.013	(.014)*	.0		.009	(.015)*	.0	
ρ_{2Q}	-.0002	(.010)*	.0		.001	(.010)*	.0	
ρ_{3Q}	-.012	(.012)*	.0		-.010	(.012)*	.0	
β	-.194	(.089)	-.176	(.087)	-.172	(.050)	-.178	(.042)
λ_T	-.644	(.196)	-.651	(.160)	-.657	(.194)	-.650	(.159)
β_T	-.755	(.222)	.729	(.205)	.726	(.222)	.738	(.204)
β_1	-.204	(.023)	-.201	(.015)	-.203	(.023)	-.201	(.015)
β_2	.213	(.024)	.213	(.015)	.211	(.024)	.213	(.015)
β_3	-.009	(.010)*	-.012	(.007)*	-.009*	(.010)*	-.012*	(.007)*
β_Q	-.582	(.198)	-.519	(.176)	-.527	(.121)	-.529	(.111)

* Not significant at the .05 level.

TABLE B.3: ANNUAL SCALE ELASTICITY ESTIMATES FROM SINGLE-OUTPUT GTL COST MODELS

YEAR	UNCONSTRAINED	HOMOTHETIC $(\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0)$	$\lambda_Q=0$	HOMOTHETIC $(\lambda_Q=0;$ $\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0)$
1952	1.08	1.06	1.07	1.06
53	1.08	1.07	1.08	1.07
54	1.09	1.08	1.09	1.08
55	1.07	1.07	1.08	1.06
56	1.09	1.09	1.10	1.09
57	1.13	1.14	1.15	1.14
58	1.23	1.24	1.25	1.23
59	1.29	1.29	1.31	1.29
1960	1.39	1.40	1.41	1.40
61	1.45	1.45	1.47	1.45
62	1.48	1.49	1.50	1.49
63	1.60	1.61	1.62	1.61
64	1.72	1.72	1.74	1.72
65	1.73	1.73	1.75	1.73
66	1.74	1.74	1.76	1.74
67	1.73	1.73	1.75	1.73
68	1.77	1.77	1.79	1.77
69	1.78	1.78	1.79	1.78
1970	1.78	1.79	1.80	1.79
71	1.78	1.80	1.80	1.80
72	1.77	1.79	1.79	1.79
73	1.75	1.77	1.77	1.77
74	1.74	1.77	1.76	1.76
75	1.72	1.76	1.75	1.76
76	1.71	1.75	1.74	1.75
77	1.69	1.75	1.73	1.74
78	1.71	1.76	1.74	1.75

TABLE B.4: TECHNOLOGY ELASTICITIES OF COST IN THE SINGLE-OUTPUT HOMOTHETIC GTL COST MODEL ($\lambda_Q=0$; $\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0$)

YEAR	TECHNOLOGY ELASTICITY OF COST
1952	-.20
53	-.26
54	-.33
55	-.44
56	-.44
57	-.37
58	-.27
59	-.23
1960	-.16
61	-.16
62	-.18
63	-.12
64	-.09
65	-.12
66	-.14
67	-.18
68	-.18
69	-.22
1970	-.24
71	-.28
72	-.33
73	-.36
74	-.39
75	-.45
76	-.49
77	-.52
78	-.53

TABLE B.5: OWN PRICE ELASTICITY OF FACTOR DEMAND IN THE SINGLE-OUTPUT HOMOTHETIC GTL COST MODEL ($\lambda_Q=0$; $\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0$)

YEAR	ε_{11} (LABOUR)	ε_{22} (CAPITAL)	ε_{33} (MATERIAL)
1952	-.30	.004	-.49
53	-.30	.01	-.49
54	-.29	.03	-.50
55	-.29	.04	-.50
56	-.29	.02	-.50
57	-.31	-.02	-.49
58	-.32	-.03	-.49
59	-.33	-.04	-.49
1960	-.33	-.05	-.49
61	-.34	-.05	-.49
62	-.34	-.05	-.49
63	-.34	-.05	-.49
64	-.34	-.05	-.49
65	-.34	-.05	-.49
66	-.34	-.05	-.49
67	-.34	-.05	-.48
68	-.34	-.04	-.48
69	-.34	-.04	-.48
1970	-.34	-.04	-.48
71	-.34	-.05	-.49
72	-.34	-.05	-.49
73	-.34	-.04	-.48
74	-.34	-.04	-.48
75	-.34	-.04	-.48
76	-.34	-.05	-.49
77	-.34	-.05	-.49
78	-.34	-.04	-.48
Standard Error:1967	.05	.02	.08

TABLE B.6: PARTIAL ELASTICITIES OF FACTOR SUBSTITUTION IN THE SINGLE-OUTPUT HOMOTHETIC GTL COST MODEL ($\lambda_Q=0$; $\rho_{1Q}=\rho_{2Q}=\rho_{3Q}=0$)

YEAR	σ_{12} (labour-capital)	σ_{13} (labour-material)	σ_{23} (capital-material)
1952	.15	1.34	-.41
53	.15	1.34	-.42
54	.12	1.31	-.43
55	.10	1.30	-.46
56	.12	1.31	-.39
57	.17	1.36	-.30
58	.17	1.38	-.22
59	.18	1.41	-.18
1960	.18	1.44	-.13
61	.18	1.43	-.11
62	.17	1.45	-.10
63	.16	1.47	-.05
64	.15	1.49	-.02
65	.15	1.48	-.007
66	.15	1.53	-.003
67	.15	1.54	-.006
68	.14	1.57	-.002
69	.14	1.58	.002
1970	.14	1.59	-.002
71	.14	1.54	.007
72	.15	1.53	.001
73	.15	1.56	-.003
74	.14	1.59	-.006
75	.15	1.57	-.007
76	.15	1.53	-.003
77	.15	1.52	-.003
78	.14	1.55	.006
Standard Error:1967	.06	.18	.15

APPENDIX C

PROXY VARIABLES FOR TECHNOLOGICAL CHANGES

In the early stages of econometric research, the proxy variables used to represent changes in Bell Canada's technology were simple ratios, such as the percentage of customer dialed long distance messages. These proxies depicted only a single aspect of developments in switching technology and did not extend for the entire observation period. In an attempt to incorporate several features of technological changes in one proxy variable, Smith and Corbo (1979) and Corbo and Breslaw (1979) introduced a variable, written as

$$D = \text{FNEW}\ [\tau\ \text{PDPH} + (1 - \tau)\ \text{ACCESS}],$$

where FNEW is one plus the percentage of main stations switched by XBAR, ESS and SP1; PDPH is the percentage of dial phones; ACCESS is the percentage of telephones with access to DDD and $\tau = Q_L/(Q_L + Q_T)$, where Q_L is local output and Q_T is toll service output.

Bell Canada used several alternative forms of this variable. The following three are referred to in the paper:

- T2: FNEW1 is defined as one plus the percentage of crossbar and electronic central offices;
- T3: FNEW2 is defined as one plus the percentage of telephones attached to crossbar and electronic central offices;
- T4: FNEW3 is defined as one plus the cumulative value of the first differences of the percentage of telephones served by the technologically most advanced switching equipment.

The variable FNEW3 requires some elaboration. It is calculated from the number of telephones attached to different types of central offices. The variable shows the increases in the percentage of telephones attached to the technologically most advanced switching equipment. The following equipments were considered to be technologically most advanced:

1952-55:	Step-by-Step
1956-60:	Step-by-Step and Crossbar
1961-67:	Crossbar
1968-78:	Crossbar and Electronic

Although this classification is arbitrary, there are considerations which suggest that it may reflect the different stages of technological development reasonably well. Step-by-step was the leading switching technology before the appearance of the first crossbar equipment in 1956. Between 1956 and 1960, the percentage of telephones attached to step-by-step equipment increased, indicating that step-by-step was still replacing manual equipment in substantial numbers, thereby representing technological progress even in the presence of crossbar.

In 1961, the percentage of telephones attached to step-by-step equipment started declining. (The number of telephones served by step-by-step continued to increase until 1973.) It would be unrealistic to say that step-by-step was still one of the leading switching technologies after 1960. Crossbar was the only representative of leading switching technology until the first electronic equipment came into existence in 1967. After 1968, crossbar and electronic switching are considered most advanced. The percentage of telephones attached to crossbar has slowed down considerably after 1974 but it is still increasing in 1978.

The first differences of percentages are cumulated from 1952 to 1977. One plus the cumulative values are computed for each year, the series is normalized around the 1967 value and the normalized series of FNEW3 is shown in Table C.1, together with T2, T3 and T4.

TABLE C.1: PROXY VARIABLES FOR TECHNOLOGICAL CHANGES IN TL/GTL COST MODELS ESTIMATED BY BELL CANADA

YEAR	T2	T3	T4	FNEW3
1952	48.39	51.05	48.57	76.18
1953	49.30	52.01	50.23	77.32
1954	50.28	53.05	52.02	78.53
1955	50.19	52.95	52.55	79.46
1956	52.39	55.51	55.54	81.12
1957	55.47	59.26	59.46	83.03
1958	61.46	65.91	66.03	84.04
1959	65.41	70.55	70.93	85.37
1960	71.85	78.00	78.86	87.17
1961	75.91	82.48	83.33	88.02
1962	79.14	84.13	84.84	89.79
1963	86.92	88.12	88.72	91.75
1964	94.70	93.66	94.11	94.00
1965	96.84	96.19	96.50	95.87
1966	99.02	98.61	98.76	97.97
1967	100.00	100.00	100.00	100.00
1968	104.12	105.06	104.88	102.41
1969	106.76	108.47	108.15	104.27
1970	108.95	110.86	110.33	106.43
1971	110.34	112.82	112.20	108.08
1972	111.76	115.78	114.99	110.27
1973	112.84	117.87	116.91	112.43
1974	114.74	121.11	119.91	115.58
1975	116.49	123.76	122.40	117.76
1976	117.55	126.41	124.85	120.31
1977	118.38	128.49	126.77	122.36
1978	121.57	132.71	130.83	123.99

Source: Bell Canada

ECONOMETRIC ESTIMATION OF SCALE ECONOMIES IN TELECOMMUNICATIONS

LAURITS R. CHRISTENSEN

DIANNE CUMMINGS

PHILIP E. SCHOECH

University of Wisconsin-Madison

The purpose of this paper is to investigate the importance of scale economies in the production of telecommunication services. The neoclassical cost function approach is applied to time series data from the Bell System, 1947-1979. Numerous variants of the translog and generalized translog cost functions are estimated, including total and variable cost functions, with several different representations of the level of technology. All versions estimated indicate significant scale economies. Most of the estimates fall within the range of 1.4 to 1.6.

I. Introduction

In most countries the provision of telecommunications is considered to be a natural monopoly. This view also prevailed in the U.S. for several decades. In the early 1970s, however, the Federal Communications Commission began to permit competitive entry in U.S. telecommunications. This decision is consistent with the view that scale economies are not substantial in telecommunications, and therefore that little if any efficiency would be sacrificed by allowing entry. In an industry as large as telecommunications a small decline in efficiency would represent a substantial welfare loss to consumers. Thus, the question of whether or not there are significant scale economies in telecommunications is vital for the formulation of appropriate public policy.

The two principal approaches to the study of scale economies are engineering cost studies and econometric estimation of the structure of cost and production. The engineering approach employs detailed specifications of technology while the econometric approach provides a broad view of the relationship among the major aggregate economic variables. Thus, the engineering approach is more suitable for studying scale economies in specific services, and the econometric approach is more suitable for assessing the importance of scale economies in the overall provision of telecommunication services. Evidence from aggregate econometric analysis does not in itself provide sufficient information on which to base policy for specific services. On the other hand, evidence on the degree of scale economies for the entire system can provide a useful check on the reasonableness of estimates from the body of engineering analyses.[1] Thus the engineering and econometric approaches are best viewed as complementary rather than competing methodologies for assessing the importance of scale economies.

In the 1970s there were several econometric studies of scale economies in the U.S. Bell System using the production function approach. Mantell (1974) and Vinod (1976) are examples of such studies, which typically estimated Bell System scale economies to be 1.2 or less. Since the mid-1970s econometric studies of the structure of production have shifted from direct estimation of the production function to estimation of the neoclassical cost function. It is generally agreed that the cost function provides a more attractive stochastic specification for regulated industries, and also provides a more direct approach to the estimation of scale economies.

Recently there have been two cost function studies of North American telecommunications published. Denny, Everson, Fuss, and Waverman (1981) have analyzed the structure of production for Bell Canada and Nadiri and Schankerman (1981) have investigated the structure of production for the U.S. Bell System. Both studies found scale economies that were much larger than those found in most of the earlier production function studies. Denny et al. reported a mean estimate of scale economies of 1.46 with an estimate of 2.23 for 1976 -- the most recent year in the sample.[2] Nadiri and Schankerman presented two alternative models with mean economies of 1.75 and 2.12[3].

The assessment of scale economies in telecommunications was not the primary objective of either the Denny et al. or Nadiri-Schankerman studies. Their estimates of scale economies deserve serious scrutiny, however, since they are far above estimates from production function studies. The purpose of this paper is to investigate the importance of scale economies in telecommunications using data from the U.S. Bell System for the period 1947-1979. Like Denny et al. and Nadiri and Schankerman, we exploit the neoclassical cost function; but our study goes beyond the previous studies by making use of a much more detailed data set and by employing a wide range of alternative specifications to assess the robustness of our findings.

We estimate numerous alternative specifications of the translog total cost function. Our primary representation of the level of technology is based on a distributed lag function of R & D expenditures by AT&T. For this representation of the level of technology, estimated scale economies cluster around 1.6. We also estimate scale economies with four alternative representation of the level of technology, all except one of which result in the same or higher estimates of scale economies. When we allow for autocorrelated disturbances, all versions but one result in scale economies greater than 1.6.

We use two approaches to investigate whether our estimates of scale economies are capturing changes in cost due to differential utilization of quasi-fixed inputs. First, we estimate the translog variable cost function with three different quasi-fixed inputs. All three specifications indicate scale economies are in the neighborhood of 1.6. Second, we partition the sample into observations reflecting relatively high and relatively low utilization of quasi-fixed factors of production. Total cost function estimates for the two subsamples result in estimated scale economies near those of the full sample. Furthermore, the hypothesis of equal scale economies in the two subsamples cannot be rejected. We conclude that our estimates of scale economies are not biased by changes in utilization of inputs over the postwar period.

We also test the sensitivity of our results to alternative functional forms for the cost function. To do so we employ the generalized translog cost function of which the translog function is a limiting case. The generalized translog total cost function uses a Box-Cox rather than logarithmic metric on output. Our finding of substantial scale economies is unaltered by the use of the generalized translog form.

It is possible to disaggregate total output and to estimate a multi-product translog or generalized translog cost function. However, it is difficult to estimate a multi-product cost function with only time series data. We attempted to estimate some translog multi-product cost functions with AT&T data. The results were consistent with substantial scale economies, but they were very poorly behaved. Fuss and Waverman and their collaborators have reported in published and unpublished papers several sets of results with multi-product cost functions, but they have all been based on the questionable assumption that Bell Canada maximizes profit with respect to outputs other than local service. Their results on scale economies have been very sensitive to minor change in data and functional form, thereby casting doubt on the strength of their findings.[4]

Our principal finding is that scale economies for telecommunications fall somewhere between those reported by the older production function studies and the recent cost function studies. Taking into account the 95% confidence bounds associated with our estimates of scale economies, our research indicates that Bell System scale economies are in the range from 1.4 to 1.8. This finding is consistent with the view that the proliferation of suppliers of telecommunications would result in a large sacrifice of efficiency due to foregone scale economies.

Comparing estimates of scale economies from engineering and econometric studies involves numerous complications. In particular, our estimates are most appropriately interpreted as an average of scale economies over the multitude of services provided by the Bell System, whereas engineering estimates generally relate to specific services. Nonetheless, it is interesting to note that Meyer et. al. (1979) have interpreted the results of engineering studies of long-distance telecommunications as indicating scale economies that fall in the range from 1.1 to 1.5, a range that overlaps substantially with our estimates.

II. The Translog Cost Function

The translog functional form was proposed by Christensen, Jorgenson, and Lau (1971, 1973), and was first used to represent a cost function allowing for the presence of scale economies by Christensen and Greene (1976). The translog form has been used in the telecommunications cost function studies of Denny, Everson, Fuss, and Waverman (1981) and Nadiri and Schankerman (1981) and in numerous other empirical applications. We write the translog cost function in the following form:

$$
\begin{aligned}
(1)\qquad \ell n\ C = {} & \alpha_0 + \alpha_Y\ \ell n\ Y + \sum_i \beta_i\ \ell n\ P_i + \omega_A\ \ell n\ A + \tfrac{1}{2}\,\delta_{YY}\ (\ell n\ Y)^2 \\
& + \tfrac{1}{2} \sum_i \sum_j \gamma_{ij}\ \ell n\ P_i\ \ell n\ P_j + \tfrac{1}{2}\,\omega_{AA}\ (\ell n\ A)^2 + \sum_i \rho_{Yi}\ \ell n\ Y\ \ell n\ P_i \\
& + \sum_i \phi_{iA}\ \ell n\ A\ \ell n\ P_i + \phi_{YA}\ \ell n\ A\ \ell n\ Y
\end{aligned}
$$

where $\gamma_{ij} = \gamma_{ji}$, C is total cost, Y is the level of output, the P_i are the prices of the inputs, and A represents the level of technology.

Any cost function must be homogeneous of degree one in input prices, which implies the following restrictions on the parameters of (1):

$$\sum_i \beta_i = 1, \sum_i \rho_{Yi} = 0, \sum_i \phi_{iA} = 0, \sum_i \gamma_{ij} = 0, \forall j.$$

Shephard's Lemma (1953) allows us to equate the cost shares (M_i) of the inputs to the logarithmic derivatives of the cost function with respect to the input prices:

$$(2) \quad M_i = \beta_i + \rho_{Yi} \ell\eta\ Y + \sum_j \gamma_{ij} \ell\eta\ P_j + \phi_{iA} \ell\eta\ A.$$

We follow standard practice in specifying classical disturbances for (1) and (2). The parameters of the cost function can thus be obtained by treating (1) and (2) as a multivariate regression and using a modification of Zellner's (1962) technique for estimation.[5]

Successful estimation of a cost function as general as (1) with time series data is rare. The number of parameters to be estimated is too large for the limited variation found in time series data.[6] Thus, rather than begin with (1) in its general form, as a point of departure we specify a restrictive version of (1). In particular we specify (1) with only the first order terms of each argument included. This specification implies a homogeneous structure of production, i.e., permits non-constant returns to scale, but does not permit variation in the degree of scale economies. In addition, this specification restricts all elasticities of substitution to be equal to unity. After estimating this relatively simple form of the cost function, we estimate numerous more general versions. This procedure permits us to investigate the sensitivity of the results to changes in specification; furthermore, it indicates the point at which the model becomes too general for successful estimation.

The degree of scale economies can be computed from any cost function as the inverse of the elasticity of total cost with respect to output: SCE = $(\partial \ln C/\partial \ln Y)^{-1}$. For the translog cost function (1) this yields:

$$SCE = (\alpha_Y + \delta_{YY} \ln Y + \phi_{YA} \ln A + \sum_i \rho_{Yi} \ln P_i)^{-1},$$

in which case SCE is a function of the levels of output, technology, and input prices. For the cases in which the parameters on the second order term in output and the interaction terms between output and the other arguments are zero, the degree of scale economies is constant at α_Y^{-1}. For our empirical work all variables have been normalized to unity in 1961, thus, α_Y^{-1} provides the estimate of SCE in 1961. Since the 1961 values of the variables are approximately equal to their sample means, α_Y^{-1} also provides a good approximation to SCE evaluated at the sample mean.

III. Data

The most difficult problem in the estimation of cost functions for telecommunications is how to represent the level of technology. Several representations have been suggested in the literature. Vinod (1976) has proposed using a distributed lag of R & D expenditures to represent the level of technology. He has constructed two variables based on this approach; the first uses R & D expenditures by AT&T, which we hereafter refer to as Bell R & D. The second Vinod index uses R & D expenditures by AT&T and Western Electric, which we hereafter refer to as Bell and Western R & D. Denny, Everson, Fuss, and Waverman (1981) claim that the introduction of direct distance dialing facilities and the changeover to modern switching equipment have been the two most important innovations in telecommunications in the postwar period. These innovations can be represented by the percentage of long distance calls directly dialed and the percentage of telephones connected to central offices with modern switching facilities.

It appears to us that the Bell R & D variable has the most justification as a representation of the level of technology for telecommunications. Thus, we adopt it as our primary specification. However, since a reasonable case can be made for the other measures as well, we consider them as alternative specifications. In addition to these four specifications of technology, which are specific to the Bell System, we also use an exponential time trend. We include a time trend since it is the variable that is used most widely in econometric studies to represent the level of technology.

Aside from the alternative representations of the level of technology, the data required to estimate the cost function have been discussed in Christensen, Cummings, and Schoech (1980). Thus, we provide here only a brief overview of the methods used to derive these data. The basic approach to the data has been to collect information at a very detailed level and then use the Törnqvist (or translog) index number procedure to aggregate up to the variable required for the cost function.[7] The Törnqvist index is superlative in the sense of Diewert (1976), and thus does not entail restrictive assumptions about the structure of production at the detailed level.

The output variable for the Bell System is based on its five principal sources of revenue: local, intrastate, and interstate services, directory advertising, and miscellaneous. These revenue categories are deflated by appropriate price indexes and then combined into a Törnqvist index of aggregate output.

In our cost function estimation we distinguish the three principal input categories of labor, capital, and intermediate or purchased materials.

Our approach to materials is the same as for output. We were able to obtain data for six major categories of materials: electricity, accounting machines, advertising, stationery and postage, services from Bell Telephone Laboratories, and "all other". These expenditure categories are deflated by appropriate price indexes and then aggregated. The steps used to obtain indexes of labor and capital input were more extensive.

Our index of labor input for the Bell System is based on hours worked by Bell System employees distinguished by occupation, experience, and age. In all, we used one hundred different categories of hours worked, which were combined into a Törnqvist index of labor input using relative wages as weights.

Our index of capital input is based on detailed data for twenty different types of owned tangible assets. For each of the twenty categories we obtained a time series of investment expenditures, which we then deflated by specific price indexes. The resulting real investment figures were used in conjunction with capital stock benchmarks and rates of replacement to obtain capital stock series via the perpetual inventory method. The benchmarks and replacement rates were based on surveys of Bell System Capital Stock for 1958, 1965, 1970, and 1978. These capital stocks, their asset prices, and rates of replacement were used along with the Bell System's cost of capital and tax information to compute capital service price weights. These weights were constructed following the methodology originally proposed by Christensen and Jorgenson (1969) and modified for regulated firms by Caves, Christensen, and Swanson (1980). We computed capital input for the Bell System as a Törnqvist index of the twenty types of owned capital, and one category of rented capital, using service price weights.[8]

The price index for capital, labor, and materials were obtained by dividing the annual expenditures for each category by the quantity indexes described above. Total cost is taken to be the sum of annual costs for capital, labor, and materials.

IV. Estimates of Scale Economies

As discussed in Section II, our point of departure is the translog cost function with only the first order terms of each argument allowed to appear with non-zero coefficients. For the level of technology we use Vinod's variable that is based on Bell R & D expenditures. The parameter estimates for this specification are presented in the first column of Table 1. Since this cost function is homogeneous in output, scale economies are equal to α_Y^{-1} for every data point. We find $\alpha_Y = .621$, with a standard error of .022. Taking the inverse of α_Y and its 95% confidence interval yields 1.61 as the estimate of SCE with 1.51 and 1.73 as lower and upper bounds.

We now generalize the basic specifications allowing for non-zero coefficients on second order terms in the arguments of the cost function. The second and third columns of Table 1 provide the parameter estimates with, respectively, second order terms in the level of technology and input prices. The term in technology is significant, as are some of the price terms. The addition of these terms has little impact on a estimated scale economies. With second order technology and price terms, SCE remains at 1.61 with lower 95% confidence bounds (hereafter simply "lower bounds") of 1.51 and 1.52, respectively.

The addition of the second order term in output (column four in Table 1) introduces δ_{YY} $\ell\eta$ Y into the SCE formula, in addition to α_Y. Thus, SCE becomes a function of the level of output, and the translog cost function becomes homothetic rather than homogeneous. The parameter δ_{YY} is significant, and the estimated SCE rises; in 1961 SCE is 1.65 and in 1979 it is 1.81. The lower bounds are 1.54 and 1.50, respectively.

The fifth through seventh columns of Table 1 present the parameter estimates resulting from entering the second order terms in pairs. The eighth column of Table 1 results from entering all these type of second order terms simultaneously. There is little change in SCE from entering second order terms in technology and prices. Nor is there much change when second order terms in output and prices are entered. However, when second order terms in output and technology are entered together (with or without the second order terms in prices), the results change markedly. SCE are found to be approximately 1.8 in 1961 and 3.2 in 1979. These estimates suggest that specifications with second order terms in both output and technology may be too general for successful estimation.

In the ninth column of Table 1 we present the general translog cost function, subject to the condition that output enters only through a linear term. Technology and price terms are allowed to enter linearly, quadratically, and with interactions between technology and prices. SCE is somewhat reduced from the previous specifications; the estimate is 1.53 with a lower bound of 1.44. However, contrary to the previous specifications, the neoclassical curvature conditions are not satisfied for all the data points. The cost function is not concave for eight years in the middle of the sample.

Changes in technology have often been modelled as augmenting individual inputs. The dual formulation of this approach is the specification that changes in technology diminish the prices of the inputs in the cost function.

Each price, P_i, in the cost function is replaced by $P_i A^{\lambda_i}$. Thus, we have the level of technology, A, entering the cost function via the three parameter λ_K, λ_L, and λ_M. This is one less than the four independent parameters associated with A in the ninth column of Table 1: ω_A ω_{AA}, ϕ_{KA}, ϕ_{LA}, ϕ_{MA} (with the restriction that $\phi_{KA} + \phi_{LA} + \phi_{MA} = 1.0$, to preserve linear homogeneity of the cost function in the input prices). The principal disadvantage of the factor augmenting specification is that the cost function becomes nonlinear in the parameters, and therefore is more difficult to estimate.

In the tenth column of Table 1 we present the factor augmented version of the homogeneous translog cost function. SCE is estimated to be 1.53 with a lower bound of 1.44.

We have found that our cost function estimates with second order terms in output tend to differ greatly from the other specifications. Nonetheless, for the sake of completeness we present two very general specifications in the eleventh and twelfth columns of Table 1. The latter is the general translog form and the former restricts the interactions between output and all other arguments to be zero. Both versions indicate SCE in 1961 to be between 1.6 and 1.8 with lower bounds of approximately 1.6. SCE in 1979 is indicated to be very large, but the bounds are quite wide. Neither of these estimated cost functions is concave over the full sample.

The SCE estimates from the specifications in Table 1 are summarized in Table 2. The estimates for 1961 and 1979 are presented, along with their lower bounds.

Table 2 provides evidence that SCE is 1.4 or higher. The lowest point estimate is 1.53 with a 95% confidence interval bounded below by 1.44. The most general forms of the cost function indicate much higher SCE, but the estimates are not stable. SCE varies little across many of the specifications. The exceptions are cases in which output is allowed to enter through second order terms and cases in which technology is allowed to enter in a very general way. The former exceptions result in very high SCE, and the latter exceptions result in SCE that is somewhat lower than the remaining specifications.

We proceed to investigate how the estimates of SCE in Table 2 are affected by using four alternative specifications of the level of technology. We use specification [1] to represent the bulk of the SCE estimates and specification [10] to represent the lower estimates. The parameter estimates for these regressions are presented in Table 3. Both of these specifications are homogeneous in output and thus have SCE that is the same for all sample points. The SCE estimates for these two specifications are presented in Table 4 for all five indexes of technology. Table 3 provides some support for our choice of the Bell R & D index as the primary representation of the level of technology. Its coefficient has a higher t-ratio than any of the other indexes of technology.

The results in Tables 3 and 4 indicate that the estimates of SCE change very little with three alternative specifications of the level of technology: (a) percentage of long distance calls directly dialed, (b) access to modern switching facilities, and (c) Bell and Western R & D expenditures. Using time to represent the level of technology results in a much higher point estimate of SCE for specification [1] and a lack of convergence in specification [10]. The problem is that time and output are very highly correlated, as can be seen by the large increase in the standard error in the coefficient when time is introduced in specification [1]. For this reason, we do not consider time further as an indicator of the level of technology.

It is possible to estimate translog cost functions containing more than one index of technology. This approach has been followed by Nadiri and Schankerman (1981), who included both time and an R & D index. We have estimated several such models, but we do not present the estimates. These models generally resulted in higher estimates of SCE, but usually one or both of the technology indexes was not significant. Time is generally not significant when entered with another index of technology. Modern Switching also continues to be insignificant when it is entered with Direct Distance Dialing. Entering Bell and Western R & D indexes separately results in higher SCE, but the Western R & D index is not significant.

For the four representations of technology in Table 4, other than time, the estimates of SCE range from 1.50 to 1.65 with lower bounds that range from 1.34 to 1.52. These results are based on an estimation method that allows for contemporaneous correlation among the disturbances of the estimating equations, but not for any serial correlation of the disturbances. We investigate the robustness of our results by permitting serial correlations in the manner discussed by Berndt and Savin (1975). We permit two distinct non-zero correlation coefficients -- one for the cost function and one for the share equations.[9] We repeat the regressions for specifications [1] and [10] for the four representations of technology. The parameter estimates are presented in Table 5, and the implied SCE are presented to Table 6.

For specification [1], SCE for all of the representations of technology are higher in Table 6 than in Table 4. Furthermore, the lower bounds are all higher than the corresponding ones in Table 4, and are in the narrow range from 1.51 to 1.64. For specification [10] the resulting estimates of SCE are substantially higher in all but one case than their counterparts in which serial correlation of the disturbances was not recognized. For access to modern switching the estimate is slightly lower.

V. Estimate of Scale Economies Allowing for Changes in the Utilization of Capital and Labor

It might be claimed that the estimates of scale economies reported in the previous section reflect variations in the utilization of factors of production in addition to the exploitation of existing scale economies. In this section we use two approaches to explore such a possibility. First, we replace the assumption of full static equilibrium with an assumption of partial static equilibrium. Second, we maintain the assumption of full static equilibrium but estimate separate cost functions for periods corresponding to peaks and troughs of the business cycle.

Brown and Christensen (1980) and Caves, Christensen, and Swanson (1981) have discussed estimation of the translog variable cost function implied by partial static equilibrium. Rather than minimization of total cost conditional on the levels of output, the behavioral assumption becomes minimization of variable cost conditional on the level of output and the level of any inputs that are quasi-fixed. Specification [10] of the variable cost function can be written:

$$\begin{aligned} \ln CV = {} & \alpha_O + \alpha_Y \ln Y + \alpha_Z \ln Z^* + \sum_i \beta_i \ln P_i^* + \tfrac{1}{2} \delta_{YY} (\ln Y)^2 \\ & + \tfrac{1}{2} \delta_{ZZ} (\ln Z^*)^2 + \delta_{YZ} \ln Y \ln Z^* + \tfrac{1}{2} \sum_i \sum_i \gamma_{ij} \ln P_i^* \ln P_j^* \\ & + \sum_i \eta_{Yi} \ln Y \ln P_i^* + \sum_i \eta_{Zi} \ln Z^* \ln P_i^* \end{aligned}$$

where $P_i^* = P_i A^{\lambda_i}$, $Z^* = Z A^{\lambda_i}$, and CV is the cost of the variable inputs. This form specializes to specification [1] if the λ_i, δ_{ij}, γ_{ij}, and ρ_{ij} are restricted to be zero and a first order term in technology is added.

We have estimated the translog variable cost function for three alternative specifications in which a portion of the Bell capital stock or labor force is treated as quasi-fixed: first, all tangible assets in which the lag between order and installation exceeds one year, including buildings, central office equipment (COE) and large private branch exchanges (LPBX); second, all employees with five or more years of experience; and third, all management employees, regardless of experience, and all non-management employees with five or more years of experience. All of these variations have been estimated using the technology variable based on the Bell R & D expenditures -- for specifications [1] and [10]. In addition specification [1] has been run with allowance for serially correlated disturbances. We were not able to achieve convergence for specification [10] with serial correlation permitted.

The parameter estimates for the nine variable cost functions are presented in Table 7. Caves, Christensen, and Swanson (1981) have shown that SCE can be computed directly from the parameters of the variable cost function as:

$$SCE = (1 - (\partial \ln CV/\partial \ln Z))/(\partial \ln CV/\partial \ln Y).$$

SCE computed from this formula are presented in Table 8. The SCE are very similar to those in Tables 4 and 6, but the confidence bounds are somewhat wider. Estimated SCE range from 1.49 to 1.75, and their lower bounds range from 1.39 to 1.52.

Our second approach to allowing for variation in the utilization of factors of production is to divide our sample into two subsamples, the first of which represents relatively high utilization, and the second of which represents relatively low utilization. We have used a combination of the U.S. business cycle and cycles in Bell System output in partitioning the sample. We include the following years in the subset reflecting relatively high utilization: 1948, 1950, 1951, 1955, 1956, 1959, 1964, 1965, 1966, 1968, 1969, 1972, 1973, 1976, 1977, 1978, and 1979.

We have estimated the translog cost function specifications [1] and [10] with our primary representation of the level of technology for the two subperiods. The number of contiguous observations in each subsample is so small that allowing for serial correlation is not appropriate. The parameter estimates are presented in Table 9 and the implied SCE in Table 10. The SCE are very similar to the corresponding estimates in Table 4. The bounds in Table 10 are somewhat wider, however, reflecting the fact that each subsample is much smaller than the full sample. Since the estimate of SCE for both specification [1] and [10] are based on a single parameter, it is straightforward to determine whether the differences between the estimates for years of high and low utilization are statistically significant. The t-ratios for the tests of equaltiy are .03 and .81, and thus we cannot reject the hypothesis of equal SCE from the subsamples.

Neither the translog variable cost function estimates, nor the split-sample estimates of this section provides any estimates of SCE that are substantially different from those of the previous section. We conclude that there is no evidence of upward bias in an estimated SCE due to a failure to control for capacity utilization.

Although the principal motivation for estimating the variable cost function was a concern over the effects of differential capacity utilization, the similarity of the results from the total and variable cost functions also permit us to infer that our results are not biased due to the Averch-Johnson (AJ) (1962) effect. If, as a result of rate of return regulation, a firm does not attempt to minimize total cost, estimates from a total cost function might be invalid. The AJ model specifies that a firm will use more than the optimal amount of capital. Whether or not the model is realistic (a matter of great controversy), the firm will attempt to minimize variable cost conditional on the level of capital. Therefore, the variable cost model will be valid even if there is an AJ effect. The fact that we obtain very similar estimates of scale economies with the total and variable cost models provides evidence that any AJ effect which might exist for AT&T is not important enough to invalidate estimates of scale economies from the total cost model.

VI. Estimates of Scale Economies Using the Generalized Translog Cost Function

Up to this point all of our estimates of scale economies have been based on the translog form of the cost function. We would not expect our estimates of scale economies to be substantially different if we had employed any other flexible functional form.[10] However, Fuss (1981) indicated that he and his collaborators had found substantial differences in estimates of scale economies for Bell Canada when the generalized translog functional form was used rather than the translog form. Thus, we have carried out some additional analysis to investigate the robustness of our findings with respect to use of an alternative flexible functional form. We employ the generalized translog form in order to address directly the question raised by Fuss.

The generalized translog form was first proposed by Caves, Christensen, and Tretheway (1980). This form differs from the translog form in that wherever output (Y) appears in the cost function, it is transformed by the Box-Cox metric rather than the natural logarithmic metric. The Box-Cox metric can be written $(Y^\theta - 1)/\theta$. As θ approaches zero, the Box-Cox metric approaches the natural logarthmic metric. Thus the translog cost function is a limiting case of the generalized translog cost function.

In Section IV we carried out most of our estimation on the two translog specifications [1] and [10]. For the generalized translog functional form [1] becomes:

$$\ell n C = \alpha_o + \alpha_Y \left(\frac{Y^\theta - 1}{\theta}\right) + \sum_i \beta_i \ell n\, P_i + W_A\, \ell n\, A,$$

and specification [10] becomes

$$\ell n C = \alpha_o + \alpha_Y \left(\frac{Y^\theta - 1}{\theta}\right) + \sum_i \beta_i\, \ell n\, P_i^* + \tfrac{1}{2} \sum_i \sum_j \gamma_{ij} \ell n Pi^* \ell n P_j^*$$

with $P_i^* = P_i A^{\lambda i}$. In both models the degree of scale economies equals

$$SCE = (\delta \ell n C/\delta \ell n Y)^{-1} = {}^{1}/(\alpha_Y Y^\theta).$$

In the translog form [1] and [10] the degree of scale economies is independent of the level of output, i.e., the forms are homogeneous in output. However, the generalized translog forms of [1] and [10] are non-homogeneous.

We estimated the generalized translog specifications [1] and [10] using the four alternative representatives of technology described in Section III. The paramter estimates are presented in Table 11. We present the corresponding estimates of scale economies, along with the lower bounds of the confidence regions, for the years 1961 and 1979 in Table 12. For specification [1], all the regressions indicate scale economies in the neighborhood of 1.65 in 1961 and 1.85 in 1979. The lower bounds are all higher, but quite similar to the corresponding specifications in Table 4. Specification [10] also shows strong evidence of significant scale economies, as can be seen from the entries in Table 12.

We also investigated the possibility that the variable cost function results are sensitive to the functional form used. The generalized translog variable cost function that is the analogue of specification [1] in Table 7 has the form

$$\ell\eta\ CV = \alpha_o + \alpha_Y(\frac{Y^\theta - 1}{\theta}) + \alpha_Z\ (\frac{Z^\theta - 1}{\theta}) + \Sigma_i \beta_i \ell\eta P_i + W_A\ \ell\eta A$$

We estimated this model using the index of Bell R & D expenditures to represent the level of technology. We also used the three representations of quasi-fixed input: quasi-fixed capital, experienced labor, and management plus experienced labor. The parameter estimates are presented in Table 13.

The scale elasticity can be estimated from the formula

$$SCE = (1-(\partial\ell\eta CV)/(\partial\ell\eta Z))/(\delta\ell\eta CV/\partial\ell\eta Y) = (1-\alpha_Z Z^\theta)/(\alpha_Y Y^\theta)$$

The scale elasticity estimates, along with the lower bounds of the confidence region, are presented in Table 14. These models show strong evidence of significant scale economies.

Due to their complexity, we were not able to successfully estimate the generalized translog form for the specification [10] version of the variable cost function. Nonetheless, it is clear from the estimates which we did complete that our results on scale economies are insensitive to which flexible functional form is used. Both the translog and generalized translog forms provide strong evidence of significant scale economies.

Footnotes

~~*A preliminary version of this paper was presented at the meetings of the Econometric Society held in Denver, Colorado, September, 1980. The authors wish to thank Douglas Caves, Thomas Cowing, and Zvi Griliches for helpful comments.~~

[1] For example, engineering estimates of large and pervasive scale economies for electric power generation have never been substantiated by econometric analyses. See Christensen and Green (1976) and Weiss (1975) for discussion.

[2] This estimate is based on the cost-output elasticities reported by Denny, Fuss, and Waverman (1981), which were attributed to Denny, Fuss, Everson, and Waverman (1981).

[3] Not enough information was provided to compute scale economies for individual years.

[4] See Fuss (1981) for references to the various versions of the Bell Canada results.

[5] The covariance matrix of the multivariate regression is singular. We overcome this problem by deleting one of the share equations at the second stage of the Zellner procedure. This provides estimates that are invariant with respect to which equation is deleted and are asymptotically equivalent to maximum likelihood estimates.

[6] Both Denny, Everson, Fuss, and Waverman (1981) and Nadiri and Schankerman (1981) reported difficulties in using the general translog specification with telecommunications data.

[7] The index can be written:

$$\ln (X_1/X_0) = \Sigma \, \overline{w}_i \, \ln (X_{1i}/X_{0i}),$$

where $\overline{w}_i$ is the arithimetic average of the expenditure weights in periods 0 and 1.

This index is one of many discussed by Fisher (1922). It has been recommended for applications by Törnqvist (1936) and subsequently by Theil (1965) and Kloek (1966). It has been used extensively by Christensen and Jorgenson (1973) and others. Diewert (1976) has shown that this index is exact for a homogeneous translog function.

[8] Bell System rented capital consists almost entirely of buildings.

[9] We were not able to attain covergence with more general specifications of the disturbance structure.

[10] Any flexible cost function can provide a second order approximation to an arbitrary structure of cost.

Table 1

Parameter Estimates for Twelve Variations of the Translog Cost Function With the Level of Technology Variable (A) Based on Bell R & D Expenditures

(Standard errors in parentheses)

Parameter	[1] First Order Terms	[2] Second Order Technology	[3] Second Order Prices	[4] Second Order Output
α_0	9.061(.003)	9.067(.004)	9.062(.003)	9.068(.003)
α_Y	.621(.022)	.620(.021)	.619(.019)	.607(.020)
β_K	.487(.007)	.483(.008)	.497(.007)	.481(.008)
β_L	.394(.007)	.397(.008)	.397(.008)	.400(.008)
β_M	.119(.003)	.119(.003)	.106(.003)	.119(.002)
ω_A	-.067(.023)	-.057(.024)	-.065(.020)	-.046(.023)
ω_{AA}	-	-.034(.017)	-	-
γ_{KK}	-	-	.057(.017)	-
γ_{LL}	-	-	-.031(.018)	-
γ_{MM}	-	-	.033(.022)	-
γ_{KL}	-	-	-.004(.017)	-
γ_{KM}	-	-	-.061(.014)	-
γ_{LM}	-	-	.027(.010)	-
δ_{YY}	-	-	-	-.040(.012)

Table 1 (continued)

Parameter	[5] Second Order Technology, Prices	[6] Second Order Prices, Output	[7] Second Order Technology, Output	[8] Second Order Technology, Prices, Output
α_0	9.066(.003)	9.067(.003)	9.057(.004)	9.056(.003)
α_Y	.617(.019)	.605(.018)	.556(.022)	.522(.020)
β_K	.496(.007)	.495(.007)	.481(.008)	.495(.007)
β_L	.398(.008)	.399(.008)	.400(.008)	.399(.008)
β_M	.106(.003)	.106(.003)	.119(.003)	.105(.003)
ω_A	-.055(.021)	-.043(.020)	-.205(.020)	-.021(.018)
ω_{AA}	-.029(.015)	-	.206(.056)	.212(.049)
γ_{KK}	.062(.018)	.065(.017)	-	.067(.017)
γ_{LL}	-.025(.019)	-.021(.019)	-	-.020(.019)
γ_{MM}	.033(.022)	.034(.021)	-	.037(.022)
γ_{KL}	-.002(.017)	-.005(.017)	-	-.005(.017)
γ_{KM}	-.060(.014)	-.060(.014)	-	-.062(.014)
γ_{LM}	-.027(.010)	.026(.010)	-	.024(.010)
δ_{YY}	-	-.036(.010)	-.180(.041)	-.180(.036)

Table 1 (continued)

	[9]	[10]	[11]	[12]
Parameter	Second Order Interactions Technology Prices	Second Order Prices, Factor Augmenting Technology	General Except Output Interactions	General
α_0	9.065(.004)	9.066(.003)	9.053(.004)	9.052(.004)
α_Y	.654(.021)	.653(.020)	.575(.023)	.604(.031)
β_K	.521(.004)	.521(.004)	.521(.004)	.528(.004)
β_L	.368(.003)	.368(.003)	.367(.335)	.362(.003)
β_M	.111(.001)	.112(.001)	.111(.002)	.110(.001)
ω_A	-.083(.023)	-	-.034(.021)	-.067(.032)
ω_{AA}	.073(.018)	-	.316(.058)	.663(.310)
γ_{KK}	.275(.018)	.273(.018)	.278(.018)	.269(.017)
γ_{LL}	.212(.018)	.206(.015)	.216(.018)	.242(.018)
γ_{MM}	-.015(.018)	-.022(.013)	-.015(.018)	.046(.023)
γ_{KL}	-.251(.016)	-.250(.015)	-.255(.016)	-.232(.015)
γ_{KM}	-.024(.013)	-.022(.007)	-.023(.007)	-.036(.007)
γ_{LM}	.038(.015)	.044(.011)	.038(.015)	-.010(.017)
δ_{YY}	-	-	-.178(.042)	.061(.222)
ϕ_{KA}	.141(.009)	-	.142(.009)	.047(.027)
ϕ_{LA}	-.144(.009)	-	-.145(.009)	-.047(.024)
ϕ_{MA}	.003(.005)	-	.003(.005)	.000(.008)
λ_K	-	.206(.040)	-	-
λ_L	-	-.264(.085)	-	-
λ_M	-	-.800(.398)	-	-
ρ_{YK}	-	-	-	.082(.023)
ρ_{YL}	-	-	-	-.098(.021)
ρ_{YM}	-	-	-	.017(.010)
ϕ_{YA}	-	-	-	-.268(.258)

Table 2

Scale Economies and Lower Bounds (95%)
Implied by Estimates in Table 1 for 1961 and 1977

	1961		1979	
Specification	Scale	Lower Bound	Scale	Lower Bound
[1]	1.611	1.507	1.611	1.507
[2]	1.613	1.510	1.613	1.510
[3]	1.615	1.523	1.615	1.523
[4]	1.648	1.545	1.811	1.501
[5]	1.620	1.528	1.620	1.528
[6]	1.654	1.559	1.800	1.639
[7]	1.799	1.668	3.246	1.427
[8]	1.811	1.690	3.288	2.315
[9]	1.528	1.437	1.528	1.437
[10]	1.532	1.443	1.532	1.443
[11]	1.738	1.612	3.029	2.088
[12]	1.656	1.501	3.958	2.511

Table 3

Paramter Estimates

Five Alternative Representations of the Level of Technology

Specification [1]: First Order Terms

Parameter	Bell R & D	Direct Distance Dialing	Modern Switching	Time	Bell & Western R & D
α_0	9.061(.003)	9.054(.003)	9.051(.004)	9.059(.004)	9.060(.004)
α_Y	.621(.022)	.608(.026)	.610(.032)	.439(.128)	.621(.041)
β_K	.487(.007)	.487(.007)	.488(.007)	.489(.007)	.488(.007)
β_L	.394(.007)	.394(.007)	.393(.007)	.392(.007)	.393(.007)
β_M	.119(.003)	.119(.003)	.119(.003)	.119(.003)	.119(.003)
ω_A	-.067(.023)	-.102(.054)	-.182(.114)	.009(.009)	-.056(.039)

Specification [10]: Second Order Prices, Factor Augmenting Technology

Parameter	Bell R & D	Direct Distance Dialing	Modern Switching	Time	Bell & Western R & D
α_0	9.066(.003)	9.059(.004)	9.063(.005)		9.069(.004)
α_Y	.653(.020)	.654(.026)	.614(.030)		.665(.040)
β_K	.521(.004)	.537(.003)	.534(.004)		.522(.004)
β_L	.368(.003)	.352(.002)	.355(.004)		.367(.003)
β_M	.112(.001)	.111(.001)	.111(.001)	Convergence	.111(.001)
γ_{KK}	.273(.018)	.270(.011)	.226(.015)	not	.262(.017)
γ_{LL}	.206(.015)	.207(.007)	.168(.014)	Achieved	.202(.014)
γ_{MM}	-.022(.013)	-.013(.007)	-.005(.002)		-.010(.006)
γ_{KL}	-.250(.015)	-.245(.008)	-.200(.014)		-.238(.016)
γ_{KM}	-.022(.006)	-.025(.005)	-.026(.004)		-.025(.005)
γ_{LM}	.044(.011)	.038(.006)	.031(.004)		-.035(.006)
λ_K	.206(.040)	.464(.085)	1.148(.183)		.209(.064)
λ_L	-.264(.085)	-.348(.202)	.329(.395)		-.124(.126)
λ_M	-.799(.400)	-2.373(1.01)	-7.299(1.46)		-1.241(.512)

Table 4

Scale Economies and Lower Bounds for Specifications [1] and [10]
Using Five Alternative Representatives of the Level of Technology

		Specifications		
Technology Variable		[1]		[10]
Base On:	SCE	Lower Bound	SCE	Lower Bound
Bell R & D Expenditures	1.611	1.501	1.532	1.443
Direct Distance Dialing	1.646	1.516	1.528	1.416
Access to Modern Switching Facilities	1.639	1.483	1.629	1.483
Time	2.278	1.438	*	*
Bell & Western R & D Expenditures	1.611	1.422	1.503	1.341

*Convergence not achieved

Table 5

Parameter Estimates Permitting Serial Correlation

Specification [1]: First Order Terms

Parameter	Bell R & D	Direct Distance Dialing	Modern Switching	Bell & Western R & D
α_0	9.078(.006)	9.077(.007)	9.075(.007)	9.079(.007)
α_Y	.593(.026)	.579(.028)	.606(.029)	.510(.051)
β_K	.463(.016)	.466(.016)	.463(.016)	.466(.017)
β_L	.387(.014)	.386(.014)	.387(.015)	.385(.015)
β_M	.150(.015)	.148(.014)	.150(.014)	.149(.014)
ω_A	-.038(.027)	-.041(.055)	-.171(.100)	.045(.047)
ρ_C	.198(.049)	.219(.047)	.233(.046)	.247(.048)
ρ_S	.931(.027)	.928(.028)	.931(.027)	.930(.028)

Specification [10]: Second Order Prices, Factor Augmenting Technology

Parameter	Bell R & D	Direct Distance Dialing	Modern Switching	Bell & Western R & D
α_0	9.083(.007)	9.082(.007)	9.080(.008)	9.081(.007)
α_Y	.536(.033)	.603(.030)	.648(.033)	.472(.052)
β_K	.522(.022)	.535(.003)	.544(.009)	.527(.021)
β_L	.350(.020)	.353(.002)	.342(.007)	.345(.019)
β_M	.129(.010)	.112(.001)	.114(.003)	.128(.010)
γ_{KK}	.115(.088)	.257(.008)	.247(.035)	.153(.094)
γ_{LL}	.013(.082)	.190(.009)	.212(.023)	.014(.104)
γ_{MM}	-.176(.111)	-.022(.014)	.026(.016)	-.208(.140)
γ_{KL}	-.152(.069)	-.235(.008)	.216(.026)	-.188(.074)
γ_{KM}	.037(.045)	-.023(.007)	-.030(.014)	.035(.046)
γ_{LM}	.139(.089)	.045(.011)	.004(.015)	.174(.113)
λ_K	.021(.121)	.494(.063)	.113(.275)	.122(.088)
λ_L	.008(.308)	-.408(.147)	-1.865(.456)	.006(.115)
λ_M	.059(.142)	-1.403(.800)	2.833(1.84)	.115(.111)
ρ_C	.244(.056)	.201(.050)	.266(.053)	.252(.050)
ρ_S	.859(.043)	1114(.092)	.652(.058)	.854(.047)

Table 6

Scale Economies and Lower Bounds for Specifications [1] and [10]
Using Five Alternative Representations of the Level of Technology
and Allowing for Serially Correlated Disturbances

	Specifications			
Technology Variable	[1]		[10]	
Based on:	SCE	Lower Bound	SCE	Lower Bound
Bell R & D Expenditures	1.685	1.552	1.865	1.661
Access to Direct Distance Dialing	1.728	1.577	1.658	1.506
Access to Modern Switching	1.651	1.508	1.543	1.399
Bell and Western R & D Expenditures	1.961	1.636	2.118	1.998

Table 7

Parameter Estimates for the Variable Cost Functions

Bell R & D Technology

	Quasi-fixed Capital		Quasi-fixed Capital Allowing for Serial Correlation
Parameter	Specification [1]	Specification [10]	Specification [1]
α_O	8.735(.005)	8.736(.004)	8.769(.012)
α_Y	.636(.074)	.917(.010)	.853(.098)
α_Z	-.113(.066)	-.435(.107)	-.416(.109)
β_K	.356(.007)	.403(.002)	.344(.027)
β_L	.476(.009)	.443(.002)	.412(.038)
β_M	.168(.005)	.154(.002)	.244(.004)
ω_A	-.131(.032)	-	-.008(.054)
γ_{KK}	-	.180(.009)	-
γ_{LL}	-	.132(.010)	-
γ_{MM}	-	-.026(.011)	-
γ_{KL}	-	-.168(.010)	-
γ_{KM}	-	-.011(.007)	-
γ_{LM}	-	.037(.008)	-
η_{YK}	-	.068(.018)	-
η_{YL}	-	.088(.018)	-
η_{YM}	-	.020(.012)	-
η_{ZK}	-	.169(.027)	-
η_{ZL}	-	-.138(.026)	-
η_{ZM}	-	-.031(.019)	-
δ_{YY}	-	-.907(.338)	-
δ_{YZ}	-	1.419(.500)	-
δ_{ZZ}	-	-2.220(.750)	-
λ_K	-	.143(.066)	-
λ_L	-	-.188(.117)	-
λ_M	-	-1.291(.501)	-
λ_Z	-	-.412(.037)	-
ρ_C	-	-	.414(.068)
ρ_S	-	-	.957(.022)

Table 7 (continued)

Parameter Estimates for the Variable Cost Functions

Bell R & D Technology

	Quasi-fixed Management & Experienced Labor		Quasi-fixed Management & Experienced Labor Allowing for Serial Correlation
arameter	Specification [1]	Specification [10]	Specification [1]
α_O	8.709(.006)	8.724(.004)	8.752(.008)
α_Y	.902(.084)	.737(.056)	.826(.081)
α_Z	-.258(.075)	-.192(.061)	-.310(.070)
β_K	.491(.008)	.729(.005)	.677(.025)
β_L	.390(.008)	.116(.004)	.073(.041)
β_M	.119(.002)	.155(.003)	.250(.045)
ω_A	-.110(.065)	-	-.041(.061)
γ_{KK}	-	.244(.030)	-
γ_{LL}	-	.048(.032)	-
γ_{MM}	-	-.005(.033)	-
γ_{KL}	-	-.148(.024)	-
γ_{KM}	-	-.095(.014)	-
γ_{LM}	-	.101(.029)	-
η_{YK}	-	-.165(.875)	-
η_{YL}	-	.196(.012)	-
η_{YM}	-	-.030(.018)	-
η_{ZK}	-	.267(.002)	-
η_{ZL}	-	-.317(.050)	-
η_{ZM}	-	.049(.030)	-
δ_{YY}	-	.009(.453)	-
δ_{YZ}	-	-.014(.732)	-
δ_{ZZ}	-	.023(1.18)	-
λ_K	-	.341(.283)	-
λ_L	-	.344(1.40)	-
λ_M	-	-2.195(3.31)	-
λ_Z	-	-.215(.721)	-
ρ_C	-	-	.256(.050)
ρ_S	-	-	.948(.027)

Table 7 (continued)

Parameter Estimates for the Variable Cost Functions

Bell R & D Technology

	Quasi-fixed Management & Experienced Labor		Quasi-fixed Management & Experienced Labor Allowing for Serial Correlation
arameter	Specification [1]	Specification [10]	Specification [1]
α_0	8.709(.006)	8.724(.004)	8.752(.008)
α_Y	.902(.084)	.737(.056)	.826(.081)
α_Z	-.258(.075)	-.192(.061)	-.310(.070)
β_K	.491(.008)	.729(.005)	.677(.025)
β_L	.390(.008)	.116(.004)	.073(.041)
β_M	.119(.002)	.155(.003)	.250(.045)
ω_A	-.110(.065)	-	-.041(.061)
γ_{KK}	-	.244(.030)	-
γ_{LL}	-	.048(.032)	-
γ_{MM}	-	-.005(.033)	-
γ_{KL}	-	-.148(.024)	-
γ_{KM}	-	-.095(.014)	-
γ_{LM}	-	.101(.029)	-
η_{YK}	-	-.165(.875)	-
η_{YL}	-	.196(.012)	-
η_{YM}	-	-.030(.018)	-
η_{ZK}	-	.267(.002)	-
η_{ZL}	-	-.317(.050)	-
η_{ZM}	-	.049(.030)	-
δ_{YY}	-	.009(.453)	-
δ_{YZ}	-	-.014(.732)	-
δ_{ZZ}	-	.023(1.18)	-
λ_K	-	.341(.283)	-
λ_L	-	.344(1.40)	-
λ_M	-	-2.195(3.31)	-
λ_Z	-	-.215(.721)	-
ρ_C	-	-	.256(.050)
ρ_S	-	-	.948(.027)

Table 8

Scale Economies for the Translog Variable Cost Function
Technology Variable: Bell R & D Expenditures

	Without Serial Correlation Adjustment Specification				With Serial Correlation Adjustment Specification	
Quasi-fixed Inputs	SCE	[1] Lower Bound	SCE	[10] Lower Bound	SCE	[1] Lower Bound
Buildings COE, LPBX	1.749	1.517	1.565	1.424	1.659	1.463
Experienced Labor (5+ years)	1.494	1.389	1.602	1.487	1.579	1.403
Management and Experienced Laobr (5+ years)	1.655	1.455	1.618	1.497	1.584	1.407

Table 9

Parameter Estimates for the Translog Total Cost Function Using Sub-samples Reflecting Differences in Utilization

Technology Variable Based on Bell R & D Expenditures

Specification [1]: First Order Terms

Parameter	High Utilization	Low Utilization
α_0	9.061(.005)	9.065(.004)
α_Y	.632(.034)	.633(.024)
β_K	.487(.010)	.486(.011)
β_L	.391(.010)	.398(.011)
β_M	.123(.004)	.115(.003)
ω_A	-.080(.036)	-.075(.027)

Specification [10]: Second Order Prices, Factor Augmenting Technology

Parameter	High Utilization	Low Utilization
α_0	9.069(.006)	9.067(.004)
α_Y	.683(.031)	.650(.025)
β_K	.520(.005)	.521(.006)
β_L	.368(.004)	.367(.005)
β_M	.112(.002)	.112(.003)
γ_{KK}	.268(.021)	.273(.030)
γ_{LL}	.205(.014)	.189(.037)
γ_{MM}	-.016(.012)	-.042(.036)
γ_{KL}	-.245(.017)	-.252(.026)
γ_{KM}	-.024(.008)	-.021(.012)
γ_{LM}	.040(.010)	.063(.034)
λ_K	.185(.051)	.194(.040)
λ_L	-.251(.125)	-.329(.140)
λ_M	-1.033(.604)	-.472(.351)

Table 10

Scale Economies for the Translog Total Cost Function
Using Sub-Samples Reflecting Differences in Utilization

Technology Variable Based on Bell R & D Expenditures

	Specification			
	[1]		[10]	
Sub-Sample	SCE	Lower Bound	SCE	Lower Bound
Years with Relatively High Utilization	1.582	1.429	1.465	1.343
Years with Relatively Low Utilization	1.580	1.469	1.538	1.427

Table 11

Parameter Estimates, Generalized Translog Cost Function Specifications [1] and [10] Using Five Alternative Representatives of the Level of Technology

Specification [1]: First Order Terms

Parameter	Bell R&D	Direct Distance Dialing	Modern Switching	Bell and Western R&D
α_0	9.068(.003)	9.064(.003)	9.063(.004)	9.067(.004)
α_Y	.605(.021)	.608(.022)	.589(.028)	.587(.040)
β_K	.481(.008)	.480(.008)	.483(.008)	.482(.008)
β_L	.400(.008)	.401(.008)	.398(.008)	.399(.008)
β_M	.119(.002)	.119(.002)	.119(.003)	.119(.002)
ω_A	-.045(.023)	-.091(.046)	-.085(.102)	-.020(.038)
θ	-.066(.021)	-.079(.019)	-.077(.021)	-.080(.025)

Specification [10]: Second Order Prices, Factor Augmenting Technology

Parameter	Bell R&D	Direct Distance Dialing	Modern Switching	Bell and Western R&D
α_0	9.067(.004)	9.060(.004)	9.061(.005)	9.029(.004)
α_Y	.651(.022)	.649(.026)	.644(.031)	.661(.042)
β_K	.520(.004)	.536(.003)	.535(.004)	.521(.004)
β_L	.368(.003)	.352(.002)	.354(.004)	.368(.003)
β_M	.112(.002)	.112(.001)	.111(.001)	.111(.001)
γ_{KK}	.272(.018)	.268(.011)	.231(.016)	.261(.017)
γ_{LL}	.202(.018)	.201(.010)	.205(.020)	.198(.019)
γ_{MM}	-.025(.018)	-.020(.013)	.052(.024)	-.014(.020)
γ_{KL}	-.249(.015)	-.245(.008)	-.192(.015)	-.236(.016)
γ_{KM}	-.022(.007)	-.023(.006)	-.040(.007)	-.024(.006)
γ_{LM}	.047(.015)	.043(.011)	-.013(.019)	.038(.016)
λ_K	.201(.035)	.424(.072)	.495(.129)	.194(.081)
λ_L	-.279(.082)	-.481(.176)	-1.620(.188)	-.170(.213)
λ_M	-.707(.429)	-1.645(.914)	1.176(.634)	-.981(1.09)
θ	-.008(.022)	-.021(.022)	-.093(.022)	-.014(.044)

Table 12

Scale Economies and Lower Bounds for Specificatons [1] and [10] of the Generalized Translog Total Cost Function, Using Five Alternative Representations of the Level of Technology

Specification [1]

Technology Variable	1961		1979	
Based On:	Scale	Lower Bound	Scale	Lower Bound
Bell R&D Expenditures	1.651	1.538	1.809	1.555
Direct Distance Dialing	1.644	1.525	1.834	1.611
Access to Modern Switching Facilities	1.698	1.534	1.887	1.590
Bell and Western R&D Expenditures	1.703	1.472	1.901	1.475

Specification [10]

Technology Variable	1961		1979	
Based on	Scale	Lower Bound	Scale	Lower Bound
Bell R&D Expenditure	1.536	1.431	1.552	1.341
Direct Distance Dialing	1.540	1.418	1.584	1.372
Access to Modern Switching	1.553	1.403	1.766	1.497
Bell and Western R&D Expenditures	1.513	1.321	1.543	1.137

Table 13

Parameter Estimates for Specification [1] of Generalized Translog Variable Cost Function
Bell R&D Technology

Parameter	Quasi-Fixed Capital	Quasi-Fixed Experienced Labor	Quasi-Fixed Management and Experienced Labor
α_0	8.739(.004)	8.741(.004)	8.738(.004)
α_Y	.786(.076)	.872(.056)	.864(.061)
α_Z	-.290(.075)	-.328(.048)	-.326(.055)
β_K	.350(.007)	.687(.009)	.692(.009)
β_L	.482(.009)	.143(.009)	.137(.010)
β_M	.168(.004)	.170(.004)	.171(.004)
ω_A	-.084(.030)	-.092(.045)	-.081(.047)
θ	-.165(.045)	-.044(.017)	-.048(.018)

Table 14

Scale Economies for Specification [1] Generalized Translog Variable Cost Function

Technology Variable: Bell R&D Expenditures

Quasi-fixed Inputs	1961		1979	
	Scale	Lower Bound	Scale	Lower Bound
Buildings COE, LPBX	1.640	1.321	1.973	1.633
Experienced Labor (5+ years)	1.523	1.326	1.610	1.373
Management and Experienced Labor (5+ years)	1.535	1.317	1.631	1.368

COMPARING THE EFFICIENCY OF FIRMS:

CANADIAN TELECOMMUNICATIONS COMPANIES

MICHAEL DENNY

Institute for Policy Analysis
University of Toronto

ALAIN DE FONTENAY

Government of Canada
Department of Communications

MANUEL WERNER

Telecommunications Consultant

Introduction*

A study of the efficiency of individual firms is seldom possible due to data restrictions. This paper reports on a unique empirical investigation of the efficiency of three telephone companies in Canada. Most of the data has been made publicly available by the telephone companies. They originally developed the data for their own separate productivity studies. Without their considerable effort this paper would not be possible.

The data base for each company is not entirely comparable. The appendix to the paper clarifies the major differences. Part of our task is to evaluate the sensitivity of our comparisons to alternative measures of the variables. This is required to investigate the possible errors arising from the limited comparability of data and to study the advantages and disadvantages of definitions of economic variables. The latter problem is broader than the veracity of the measured variables. Telecommunications' firms offer a wide variety of services through their networks. There are alternative sensible definitions of economic variables which will alter the magnitude and perhaps ranking of the firms' efficiency. While not wishing to obscure the results, we believe that the complexity introduced by the alternatives provides a much better understanding of the detailed changes of efficiency within and across firms.

Given a set of data on the prices and quantities of inputs and outputs, the methods we use to compare efficiency have been discussed elsewhere by us (Denny, de Fontenay and Werner (1980a,b), Denny and Fuss (1980a,b) and by Caves, Christensen and Diewert (1980)). In this paper, we will apply these methods without extensive discussion due to space limitations.

* John Veitch provided excellent research assistance.

/

An Introduction to the Companies

In this section, we want to provide a descriptive analysis of the three companies. Bell Canada (Bell), Alberta Government Telephone (AGT) and British Columbia Telephone (BC Tel.) are the largest common carriers in Canada and provide a very wide range of telecommunications services within their geographic service area. Bell and BC Tel. are private companies whose tarrifs and rates of return are federally regulated. AGT is a crown corporation, i.e., a public enterprise, in the Province of Alberta.

In 1978, AGT, Bell and BC Tel. provided about 75% of the dollar value of domestic telecommunications services in Canada. In Table I, the structure of revenue and costs for these companies in 1978 is presented. Bell is by far the largest company with revenues that are roughly five times larger than either of the other two firms.

The operating revenue of the three firms is derived from local, long distance and other services. In 1978, the proportions of revenue derived from these three broad service categories were quite different. Bell received over one-half of its revenue from local services while AGT received less than one third. BC Tel. generated about 43% of its revenue from local services. AGT provides long distance services for the Edmonton Telephone Co. The latter firm provides local services for one of the largest urban areas in Alberta. If one combined AGT with Edmonton Tel.

Table I

Operating Revenues and Costs in 1978
(millions of dollars, percentages in brackets)

		AGT	BELL	BC Tel.
1.	Operating Revenue	444	2497	551
2.	Local	138 (31)*	1263 (51)	242 (43)
3.	Long Distance	292 (66)	1153 (46)	319 (57)
4.	Other	17 (4)	94 (4)	-2.3 (0)
5.	Operating Cost	339	1785	393
6.	Maintenance	87 (26)**	420 (23)	109 (28)
7.	Depreciation	125 (37)	474 (27)	113 (29)
8.	Traffic	24 (7)	127 (7)	40 (10)
9.	Marketing	29 (9)	141 (8)	46 (12)
10.	Other	64 (19)	481 (27)	58 (15)
11.	Non-Income Taxes	9 (3)	141 (8)	28 (7)

*percentage of operating revenue

**percentage of operating costs

Source: Statistiques Financières sur les Sociétés Exploitants de Télécommunications du Canada.

the revenue shares would be very similar to those of BC Tel. Consequently, it may be suggested that AGT's high long distance revenue share is partially due to the existence of a large urban local service company within AGT's territory.

The 1978 operating costs for the companies have also been broken down in Table I. For all companies maintenance and depreciation are over 50% of total operating costs. Bell appears to have a lower share of costs devoted to maintenance than the other companies. AGT has a very high depreciation cost share. Bell has tended to have a larger share of other costs than BC Tel. and AGT.

The static situation portrayed in Table I may disguise rapid shifts in the importance of the revenue and cost components due to growth through time. To characterize shifts through time, Table II shows the 1978 values of the revenue and cost component as indexes with base year 1972. Revenue growth has been much faster for AGT than for Bell and BC Tel. The growth in long distance revenue has been faster than local revenue growth in Bell and BC Tel. but in Alberta, the rapid population growth has provided a very rapid growth even in the local service revenue component.

Total operating costs have grown proportionately with revenue for AGT but have exceeded revenue growth in Bell and BC Tel. For all companies traffic costs have grown more slowly than total costs. For AGT, the growth in depreciation and maintenance costs has been higher and in non-income taxes, lower than total costs. Bell's other costs grew much more while depreciation and marketing grew less than the firm's total cost. Marketing and non-income tax costs grew faster than average and maintenance costs grew more slowly in BC Tel. While there is some diversity in the

Table II

1978 Indexes of Revenues Operating and Operating Costs, 1972=100

	AGT	BELL	BC Tel.
Local Revenue	319	201	227
Long Distance Revenue	315	248	278
Total Revenue	314	222	242
Total Cost	314	233	246
Maintenance	329	217	222
Depreciation	342	208	260
Traffic	217	192	201
Marketing	311	203	315
Other Costs	309	310	236
Non-Income Taxes	248	261	321

Source: See Table I.

revenue and cost growth and shares, it is not sensible to conclude anything about efficiency from these data. They will provide some questions which we will attempt to explore in more depth later in the paper.

A further simple comparison of these companies can be based on the number of telephones per employee. Very roughly this measures the magnitude of the network served by each employee. The companies differ enormously in the value of this measure (see Table III). Of the three major companies, Bell has the largest number of telephones per employee followed by BC Tel. and AGT. There are some sharp fluctuations in the annual series and no dominant trends.

What do these differences signify? The AGT numbers are extremely low and this appears to be a function of the low average density of the AGT area served. Edmonton Telephones is included in Table III to provide a contrast. Their urban network has a very nigh number of telephones per employee. If we combine Edmonton Tel. with AGT, the results are very similar to those for BC Tel. If this interpretation is correct the high numbers for Bell may only signify a more densely packed network.

Table III

Telephones per Employee

	BC Tel.	AGT	BELL	EDMON. TEL.
1972	109	85	166	240
1973	98	87	165	250
1974	99	84	162	230
1975	112	82	176	222
1976	112	86	173	220
1977	121	90	171	220
1978	121	95	168	245

Source: See Table I.

/

Productivity as Measured by the Companies

All three companies have produced productivity measures and for reference purposes, we have included some of their estimates here. In Table IV, the estimates of productivity produced by the company are shown. BC Tel. and Bell Canada have calculated estimates of total factor productivity growth rates. From 1972-79, Bell has had an average rate of growth of TFP of 3.1% compared to the lower BC Tel. average of 2.6%. Given the differences in the methods used, the Bell - BC Tel. results may be closer than their numbers indicate.

AGT and Bell produce estimates of value-added productivity. AGT's productivity has grown at 7.2% a year which is substantially higher than Bell's average of 4.0%. Without any serious investigation of methodology, the ranking using these measures would be AGT first and Bell and BC Tel. tied. There is no doubt that these are very high rates of productivity growth relative to other industries. Our task is to evaluate why these results were achieved and to provide a more detailed underpinning for these results.

Measured productivity growth is often correlated with output growth. This is expected since accurate measures of utilization of quasi-fixed inputs is seldom possible. In periods of slow output growth, productivity growth is low since the input measurement incorrectly overestimates utilization which falls as firms maintain input levels over fluctuations in demand growth. This may be a more serious problem in telecommunications due to the high weight of relatively fixed capital and the labour required to maintain it.

Table IV

Company Measures of Productivity Growth

	Total Factor Productivity		Value-Added Productivity	
	BC Tel.	BELL	BELL	AGT
1967	—	5.7	6.6	—
1968	—	3.9	4.5	6.9
1969	—	2.9	7.4	6.8
1970	—	3.5	4.2	5.5
1971	—	-1.0	-1.0	4.7
1972	0.3	3.8	4.5	11.5
1973	2.8	4.8	5.7	9.0
1974	5.7	4.7	5.6	14.2
1975	5.9	6.9	8.2	9.9
1976	4.7	1.0	1.2	0.7
1977	-3.6	0.7	0.8	7.2
1978	2.5	2.0	2.5	2.7
1979	2.4	1.3	1.5	—

Source: See data appendix.

In Table V, the companies' output growth rates are shown. First one can see that AGT has had a very high rate of output growth underlying their high rates of productivity growth. BC Tel.'s output grew at 10.22% compared to Bell's output growth of 8.8% from 1972-79. These are less than 60% of AGT's output growth rate. For all companies relatively high average rates of output growth have accompanied relatively high rates of growth of productivity.

Table V

Company Measures of Output Growth Rates

	BC Tel.	BELL	AGT
1967	—	9.1	—
1968	—	9.1	10.5
1969	—	10.4	13.7
1970	—	9.5	12.1
1971	—	5.6	10.6
1972	9.0	8.1	15.7
1973	11.0	10.7	13.9
1974	14.3	11.0	20.1
1975	10.3	11.0	19.0
1976	9.2	7.6	12.2
1977	6.3	6.9	13.6
1978	9.8	8.7	19.2
1979	11.7	6.3	—

Source: See data appendix.

Labour Productivity and Labour Efficiency Levels

To begin our comparison, we have estimatedlabour productivity and compared the companies on their levels of labour productivity. Output is the aggregate of the output disaggregation provided by the firms and discussed in the appendix. For reasons of comparability, labour is measured as unweighted man-hours of labour worked in each company.

In Table VI, indexes of labour productivity for AGT, BC Tel. and Bell are shown. Labour productivity in AGT and BC Tel. has grown at approximately 8% a year since 1972 compared to about 4.5% in Bell. Prior to 1972, labour productivity was growing at an annual rate above 10% at AGT and 7.7% at Bell Canada.

Output growth was higher at BC Tel. than at Bell after 1972. Labour input must have grown faster at Bell than at BC Tel. during this period in order to convert BC Tel.'s 2% advantage in output growth into a 3½% difference in labour productivity growth. AGT had the fastest rate of growth of output after 1972 but this was not translated into a higher labour productivity growth relative to BC Tel. Given the rates of growth of output, BC Tel. has managed a superior performance relative to Bell and AGT in achieving labour productivity growth.

The levels of labour productivity are presented in Table VII. Bell Canada's labour productivity level is normalized to 100 in 1972 and the data for the other companies is relative to this normalization. Bell has had the highest level of labour productivity. The other two companies have reduced the gap. After 1975, the change in the relative levels has slowed down as each company has had increasing difficulty in raising its labour productivity level.

Table VI

Labour Productivity
(1972 = 100.0)

	AGT	BCT	BELL
1967	61.7	—	66.3
1968	70.7	—	74.4
1969	76.7	—	80.8
1970	81.4	—	86.2
1971	88.2	—	92.5
1972	100.0	100.0	100.0
1973	107.2	104.2	105.4
1974	121.8	111.9	109.7
1975	143.8	131.4	122.3
1976	149.3	150.8	125.5
1977	164.1	159.9	129.6
1978	159.3	157.1	131.7
1979	—	149.2	133.9

Source: See data appendix.

Table VII

Levels of Labour Productivity
(Index, Bell 1972 = 100.0)

	BC Tel.	AGT	BELL
1967	--	43.6	66.2
1968	--	50.0	74.6
1969	--	54.3	80.6
1970	--	57.5	86.2
1971	--	62.5	92.6
1972	82.0	70.9	100.0
1973	84.7	75.6	105.2
1974	91.7	86.2	109.9
1975	107.2	102.0	121.9
1976	123.4	105.3	125.0
1977	129.8	116.3	129.8
1978	128.2	112.3	131.6
1979	121.9	--	133.3

Total Factor Productivity

We will measure total factor productivity for AGT, Bell and BC. Tel. using a common methodology and data which is partially standardized. Define the rate of growth of productivity,

$$\dot{TFP} = \dot{Q} - \dot{F}$$

where the aggregate output growth rate $\dot{Q}$ is defined by,

$$\dot{Q} = \sum_j r_j \dot{q}_j$$

and the aggregate input growth rate, $\dot{F}$ is defined by,

$$\dot{F} = \sum_j s_i \dot{x}_i \quad .$$

The disaggregate output $(\dot{q}_j)$ and input $(\dot{x}_i)$ growth rates are weighted by the revenue (r_j) and cost (s_i) shares respectively. This standardizes the methodology for the three companies.

The data are partially standardized by the choice of input variables. For each company, labour input is measured as man-hours worked without any adjustment for skill levels. Capital is measured as the gross capital stock which is an aggregate of detailed physical assets. Material inputs are not completely comparable but this is not believed to be a problem. Finally, the assumption is made that the value of capital services can be measured as a residual component in total realized costs.

Table VIII

Annual Rates of Growth of TFP

	BC Tel.	AGT	BT
1967	—	—	5.9
1968	—	5.3	4.3
1969	—	5.5	2.9
1970	—	4.6	3.7
1971	—	4.2	-0.5
1972	—	9.3	3.7
1973	2.9	7.7	4.7
1974	5.9	11.9	4.4
1975	6.0	8.3	6.9
1976	4.4	3.3	1.0
1977	-2.2	6.6	0.7
1978	3.0	2.0	2.3
1979	2.5	—	2.2

Source: See data appendix.

Table IX

TFP Indexes
(1972 = 100)

	BC Tel.	AGT	BT
1967	—	74.9	86.8
1968	—	78.9	90.6
1969	—	83.4	93.3
1970	—	87.3	96.8
1971	—	91.1	96.3
1972	100.0	100.0	100.0
1973	102.9	108.0	104.8
1974	109.1	121.7	109.5
1975	115.9	132.3	117.3
1976	121.0	132.8	118.5
1977	118.4	141.8	119.4
1978	122.0	144.8	122.2
1979	125.1	—	124.9

Source: See data appendix.

For the three companies, the rates of growth of total factor productivity are shown in Table VIII and a productivity index (1972 = 100) appears in Table IX. The standardization of methods and data does not alter our earlier comments based on the companies published results. AGT has had a faster rate of growth of TFP than Bell and BC Tel. during any time period when comparable data is available. From 1972-78, AGT's productivity grew at an average annual rate of 6.6% compared to a rate of 3.9% for Bell and for BC Tel.

Recall that AGT and BC Tel. had almost identical rates of growth of labour productivity. The TFP results indicate that BC Tel. achieved the labour productivity results through faster rates of growth of the capital-labour and the materials-labour ratio relative to AGT. The latter company was more successful at achieving high rates of labour productivity growth via high rates of TFP growth.

Bell had a substantially lower rate of growth of labour productivity than BC Tel. but TFP grew at least as quickly. Relative to Bell as well as AGT, BC Tel. must have had a faster rate of growth of capital and materials to labour intensities in order to achieve the results portrayed above.

Relative Efficiency

Relative efficiency will be measured using the methodology originally proposed by Jorgenson and Nishimizu (1978). This methodology has been developed more extensively by Denny and Fuss (1980a, b, 1981), Caves, Christensen and Diewert (1980) and Denny, de Fontenay and Werner (1980a, b,). The discussion of these methods will be relatively brief since they are more elaborately developed in the cited papers.

One can provide an intuitive interpretation of the method. It would be straightforward to compare the efficiency of the firms if we observed them using the same vector of inputs. Then, the relative output level would measure the relative efficiency levels. As Caves, Christensen and Diewert (1980) have shown, the relative efficiency measure we use has the following interpretation. Our relative efficiency measure equals the average of the relative efficiency levels of the firms measured as the relative output levels at each firm's input level. That is, it is equal to the average of the relative output levels when both firms use the observed input levels of one firm. A similar interpretation may be given to the cost efficiency measure. These interpretations imply that the differences in the prices and the quantities of inputs and outputs across firms are accounted for in the relative efficiency measure.

Using the data underlying our calculations of total factor productivity, an initial comparison of the firms' relative levels of efficiency was made. Define the relative total factor productivity level, of firm k relative to firm h, E_{kh}

$$\log E_{kh} = \log (Q_k/Q_h) - \tfrac{1}{2} \sum_i (s_{ik} + s_{ih}) \log (X_{ik}/X_{ih}) \quad ,$$

where s_{ik} is the cost share of factor i in firm k and X_{ik} is the equivalent quantity.

From the cost function, one may define a relative cost efficiency level, CE_{kh}

$$\log CE_{kh} = \log(C_k/C_h) - \tfrac{1}{2} \sum_i (x_{ik} + x_{ih}) \log (w_{ik}/w_{ih}) - \log (Q_k/Q_h) \quad ,$$

where C_k is the total cost and w_{ik} the price of input i in firm k .

Tables X and XI present the results, E_{kh} and CE_{kh} , of measuring both of these relative efficiency measures for the three companies. Consider the results of comparing Bell and AGT in Table X. In 1967 Bell's relative TFP level was 124.8 compared to AGT's 100. Alternatively, one may state that the quantity of output produced by Bell was approximately 25% greater than that produced by AGT after accounting for differences in input quantities. For the companies to be equally efficient, the E value for Bell would have to bé 100.

The results are roughly equivalent when measured from the cost side. Bell's cost efficiency in 1967 was 80.2 relative to AGT's 100. Bell's costs were only 80.2% of AGT's after accounting for differences in input prices and output levels.

Through time AGT has eliminated most of the relative efficiency gap. In 1978 there is almost no difference in the relative efficiency level. In our explorations below we will try and indicate what led to this sharp improvement in AGT's relative efficiency.

In Table XI, AGT and Bell are compared to BC Tel. for the years 1972-78. In 1972, BC Tel. and Bell had approximately equal efficiency and BC Tel. was 10% more efficient than AGT. Since BC Tel. and Bell had equal average productivity growth during this period there is no substantial change in their relative efficiency levels during the '70's. Since AGT had a very rapid growth in TFP relative to the other companies, the initial efficiency disadvantage of AGT relative to BC Tel. had been sharply reversed. AGT began in 1972 with a 10% cost disadvantage and finished with a 7% cost advantage.

Table X

Relative Efficiency of Bell Compared to AGT

	Productivity		Cost Efficiency
	BELL	AGT	BELL
1967	124,8	100	80.2
1968	123.9	100	80.7
1969	120.9	100	82.7
1970	120.4	100	83.1
1971	115.6	100	86.5
1972	109.7	100	91.2
1973	106.4	100	93.9
1974	98.8	100	101.2
1975	98.3	100	101.7
1976	98.9	100	101.1
1977	93.3	100	107.1
1978	93.4	100	107.1

Table XI

Relative Efficiency of AGT and Bell Compared to BCT

	Productivity			Cost Efficiency	
	AGT	BELL	BC Tel.	AGT	BELL
1972	89.6	98.8	100	111.7	101.2
1973	94.1	100.7	100	106.3	99.4
1974	100.0	99.5	100	100	100.5
1975	102.4	101.0	100	97.6	99.0
1976	98.6	98.1	100	101.4	102.0
1977	108.2	101.2	100	92.4	98.8
1978	107.5	100.5	100	93.0	99.4

Interpreting the Results

Our investigation is limited by the data that we have available publicly. The results suggest that in 1978 Bell and BC Tel. used more real resources to produce a given output level than AGT. To clarify this possibility, we will study the use of each factor and the production of outputs for the three companies. To begin, consider the indexes of the input-output ratios for each factor and company presented in Table XII. The indexes are normalized to 100 for Bell Canada in 1972.

For Bell Canada, the labour to output ratio has declined throughout the period. However the decline was more rapid prior to 1972 than after. BC Tel. had a much larger labour-output coefficient in 1972 but the ratio declined more quickly for BC Tel. than Bell after 1972. There was still a slightly lower labour coefficient in Bell in 1979. AGT had a very high labour coefficient relative to Bell in 1967 but this coefficient has declined more rapidly for AGT than Bell. Most of the large difference had disappeared by 1979. For the input labour, both BC Tel. and particularly AGT have done better than Bell. Notice that this ranking corresponds to the ranking of the output growth rates among the companies. To what extent that output measures are biasing the results will be investigated below.

The capital-output has fallen for Bell but the temporal pattern is reversed. Prior to 1970 the capital coefficient fell very slowly and after 1972 its rate of decline increased. The rate of decline was always much slower than the decline in the labour coefficient. The capital-labour ratio has increased in Bell throughout this period.

In 1972, the capital coefficient of BC Tel. was lower than at AGT or Bell. The very slow reduction in the BC Tel. capital coefficient has

Table XII

Input-Output Ratios
Indexes: BELL 1972 = 1.00

	Labour			Capital			Materials		
	AGT	BCT	BELL	AGT	BCT	BELL	AGT	BCT	BELL
1967	2.29	-	1.51	1.25	-	1.06	0.92	-	0.97
1968	2.00	-	1.34	1.25	-	1.05	0.91	-	0.94
1969	1.84	-	1.24	1.19	-	1.02	0.87	-	1.01
1970	1.74	-	1.16	1.15	-	1.00	0.84	-	0.94
1971	1.60	-	1.08	1.13	-	1.01	0.81	-	1.05
1972	1.41	1.22	1.00	1.06	.92	1.00	0.72	0.81	1.00
1973	1.32	1.18	0.95	0.98	.90	0.96	0.64	0.79	0.96
1974	1.16	1.09	0.91	0.87	.87	0.91	0.58	0.70	0.91
1975	0.98	0.93	0.82	0.83	.88	0.88	0.60	0.65	0.81
1976	0.95	0.81	0.80	0.82	.88	0.88	0.66	0.66	0.82
1977	0.86	0.77	0.77	0.80	.90	0.87	0.57	0.84	0.86
1978	0.89	0.78	0.76	0.74	.88	0.84	0.61	0.72	0.86
1979	-	0.82	0.75	-	.83	0.82	-	0.66	0.82

eliminated the gap relative to Bell and AGT at the end of the period.

At AGT, the capital coefficient has fallen throughout the period at a rate faster than either of the other companies. The large (50%) gap relative to Bell that existed in 1967 has been substantially reduced by 1978. While the capital to labour ratio increased sharply prior to 1972, its growth has been much slower absolutely and relative to the other companies after 1972.

For materials the pattern is different since at the beginning of the period Bell did not have a substantially lower materals coefficient. Instead it was modestly higher. At Bell, the materials coefficient has fallen by less than the other coefficients. The other two companies have maintained their lower materials' coefficient throughout the period and after 1972 there has been little change in the relative coefficients. Prior to 1972 AGT's materials coefficients did fall more than Bell's coefficient. The advantage held by BC Tel. and AGT over Bell does not result in a very large impact on the comparison for two reasons. Materials are the least important input due to their smaller cost share and the differences across companies is smaller than the differences in the other two inputs.

These results suggest that in relative terms AGT has improved its efficiency level through improved utilization of labour particularly. The same pattern is observed for BC Tel. AGT has also improved its capital utilization but this has not been as spectacular.

Alternatives

There is an extensive literature on international comparisons of productivity. Kravis (1976) has produced a fine survey and he has been one of the major researchers in the large United Nations study reported in Kravis et. al. (1975, 1978). We have used their methodology adapted to our situation and the results do not change. The major results are identical and even the numerical magnitudes are very close. This has convinced us that our results are not sensitive to quite large changes in the methodology used to measure relative efficiency.

We have also attempted to assess the impact of alternative measures of the price and flow of capital services. This required the development of relatively simple user costs of capital. These replaced the implicit user costs inherent in our comparison discussed above. Once again, the results did not change. However, we are developing a more detailed specification of the cost of capital for each company which will provide more accurate estimates.

Finally, we should note that the relative efficiency differentials reported in this paper include all the effects of regulation, non-competitive behavior and scale economies. Any separation of the relative efficiency levels into these types of causes requires econometric analysis. The data series were not long enough to permit this type of study.

Conclusions and a Warning

This is an attempt to compare the efficiency of the telephone companies using the aggregate publicly available data. Our major conclusion is that AGT has made major strides in improving its relative efficiency level compared to BC Tel. and Bell Canada. The latter two companies have had roughly equivalent efficiency levels with no major changes in their relative efficiency.

We do not believe that our results will change until we have better data. However, we do expect that some changes will occur as we are able to incorporate more disaggregate and accurate data. Consequently, we would recommend that these results be viewed as the best that currently exist but ones that may change with the further work that we are currently doing.

It must be remembered that neither profit or efficiency levels explain themselves. One may know that efficiency or profits are high or low but it is a more extended task to ascertain why these results occurred. One should not use the results given here to imply any particular line of causation since we have not developed any causes for the differences in relative efficiency.

Data Appendix

The comparisons that have been made are based on the public data bases of the three companies. In a small but crucial number of incidents the companies have provided extra data which was very helpful. The purpose of this section is to identify the exact public data series which were used.

For Bell Canada, the data were taken from the most recent productivity submission to the CRTC:

> Bell Canada, Information Requested by National Anti-Poverty Organization, March 30, 1981, Bell (NAPO) 30 Mar. 81-612, CRTC.

For BC Telephone the data were taken from the submission to the CRTC:

> BC Telephone, Total Factor Productivity Study: Data Description and Methodology, by J.T.M. Lee, BC Tel. (NAPO) 80-08-01-406 CRTC.

For AGT, data in current dollars was supplied by the company and the corresponding constant dollar data appear in the CRTC submission by AGT, Saskatchewan Telecommunications and Manitoba Telephone Systems in the CNCP-Bell Canada inter-connect case:

> Some Economic Aspects of Interconnection, Evidence in Chief, H. Harries, economic witness.

BELL CANADA

Labour

- the quantity equals the unweighted man-hours (MH) (unadjusted man-hours from Table 6 of NAPO 30 Mar. 81).

- the price index, PL, is generated by dividing total labour compensation (TLE) (Table 6 NAPO 30 Mar. 81) by unadjusted man-hours.

- PL = TLE/MH $

Capital

- total average gross stock of physical capital in current $ divided by constant $ series (Table 7 NAPO 30 Mar. 81) yields the asset price series. This asset price series was re-normalized in 1972 and the re-normalized price was divided into current $ total average gross stock of capital to yield a constant dollar gross capital series in 1972 $.

The value of capital services was generated residually by subtracting total labour compensation (Table 6 NAPO) and current $ cost of materials (Table 3 NAPO) from Total Revenue (Table 1 NAPO 30 Mar. 81)

$$VK = TR - PM^{*} M - PL * L$$

The service price of capital was arrived at by dividing the 1972 constant $ gross capital series into the value of capital services.

Materials

- current $ cost of materials, services, rents and supplies divided by constant $ cost of materials, etc. (from Table 3 NAPO 30 Mar. 81) to arrive at a price index. This price series is re-normalized in 1972. The re-normalized series is divided into the current $ cost of materials to provide a constant $ material series.

Output

- the output quantity is a divisia index with the output price = 1.0 in 1972. The components in the divisia index are the prices and quantities of local service, message toll, other toll, directory advertising and miscellaneous. Current and constant $ amounts for these categories appear in Tables 1 and 2 of NAPO 30 Mar. 81. The price series for each classification were found by dividing current $ series by the corresponding constant $ series.

B.C. TELEPHONE

Labour

- Table A-13 of (BC tel. NAPO 80-08-01-406) provides expensed labour hours and expensed wages, benefits and taxes for the following classifications; management, clerical operators, occupational, engineers, salesmen, service rep., technicians and draftsmen. The quantity of labour is the simple, unweighted sum of the expensed labour hours of all these categories. The price of labour was found by dividing this quantity of labour into the unweighted sum of the expensed wages of all the categories.

$$PL = \frac{\sum_i \text{wages}_i}{\sum_i \text{labour hours}_i}$$

Capital

- the value of capital services was found as the sum of the financial charges (Total line in Table A-4), depreciation (Total line in Table A-5), property tax (Total line in Table A-6) for Okanagan Tel. and the financial expense (Total line in Table A-7), depreciation expense (Total line in Table A-8) and property taxes (Total line in Table A-9) for B.C. Tel.

The capital series was found as the reproduction cost of capital in Table A-11, adjusted to 1972 $.

The price of capital services was generated by dividing the value of capital services series by the capital series.

Materials

- the value of materials is generated residually. It is found by subtracting total expensed wages (see above) and the value of capital services (see above) from total revenue (see above).

This value of materials series is deflated by a re-normalized (1972) materials price index equal to the Stats Can GNE deflator to yield a constant 1972 $ series for materials.

Output

- the output price and quantity series is a divisia index (price = 1.0 in 1972) of the disaggregated output categories given in Tables A-1 and A-2. The quantity series is given in Table A-2 while the corresponding revenues are given in Table A-1. A price series is generated for each category by dividing the quantity series into the revenue series.

ALBERTA GOVERNMENT TELEPHONES

Labour

- current $ value of labour (from Harries testimony)is divided by the man-hour series(Interconnection Evidence, App. 4, Table 1) to arrive at a price series for labour. No normalization is performed on these series.

Capital

- the value of capital serices in current $ (from Harries Testimony) is divided by constant 1972 $ average gross capital series to yield a price of capital services. This series is constructed by dividing the current $ gross capital series (Harries testimony)by the constant 1971 $ gross capital series (Interconnection Evidence) which yields an asset price series. The asset price series is re-normalized in 1972 and then divided into current $ gross capital to arrive at the constant 1972 $ gross capital series. The price of capital services is arrived at in this manner.

Materials

- the current dollar value of materials (in Harries letter of Dec.4, 1980) is divided by the constant 1971 $ value of materials (provided in Interconnection Evidence Appendix 4, Table 1) to arrive at a price series. This price series is re-normalized in 1972 and a constant $ material series is found by dividing current $ value materials by the re-normalized price series.

Output

- the output quantity series is produced by dividing gross revenue in current $ (Harries letter) by gross revenue in constant 1971 $ (Inter. Ev.) to yield an output price series. The output price is re-normalized in 1972 then divided into current $ gross revenue to yield a constant 1972 $ output series.

Non-Comparable Data

It was not possible to change the public data bases to eliminate some difficulties. Two areas require further improvement. First, BC Tel. measures aggregate output as the Divisia index of disaggregate quantities. The other two companies use a finer disaggregation of output prices to calculate an aggregate output price index and an implicit quantity index. Due to the more aggregate BC Tel. procedure, the growth rate of BC Tel. output is undoubtedly underestimated. We do not know the magnitude but we can be certain of the direction. Second, the differences in the relative output price levels are underestimated for AGT. This affects the level of AGT output and tends to depress it. Consequently, we have undoubtedly underestimated the level but not the growth rate of AGT's output. Correcting this will reduce AGT's disadvantage in relative efficiency during the early years.

THE BELL CANADA PRODUCTIVITY STUDY

F. KISS

Bell Canada

1. Introduction

The Bell Canada productivity study was conceived in the mid–1960's as a consequence of management's realization that, in addition to the multitude of operational efficiency measures that had existed for several decades and were used essentially by lower and middle management as a tool of control and evaluation in their everyday work, all-inclusive aggregate economic measures of performance were needed. The main purpose of total factor productivity measures was the broad evaluation of overall productive performance for executive and regulatory use.

The productivity study has increased in complexity as measurement methods were gradually refined and the analysis of productivity gains and their impact on the company's operations was expanded. At the present time, the Bell Canada productivity study consists of five chapters; viz.,

1. measurement of productivity gains,
2. analysis of productivity gains,
3. analysis of the impact of productivity gains,
4. productivity gains in the company budget,
5. productivity gain comparisons.

Measurement methodology is undergoing continuous improvement as refinements are made in output and input data and restrictive assumptions about the underlying production model are relaxed. The analysis of productivity gains utilizes aggregate econometric cost models of Bell Canada in an effort to establish the approximate effect on productivity gains of technological changes, economies of scale and other factors. The impact of productivity gains is analyzed with respect to the net income of the company and also with respect to output prices. The latter is capable of giving an approximate measure of the company's ability to absorb inflationary input price increases through improvements in the efficiency of production. Productivity gains implicit in the corporate budget have been computed for the last seven years. The traditional budget information is transformed into economic variables which are used to calculate productivity gains.

Bell's productivity gains are compared to those of various segments of the Canadian and U.S. economy. Some initial analysis has been done on the causes of observed differences in productivity gains. Work in this area will resume after the completion of the Department of Communications' research project on comparative efficiency in Canadian telecommunications.

This paper has two main objectives. First, it discusses the issues of productivity measurement and analysis that telephone companies must face when developing a study of their productivity improvements. (It does not elaborate on the impact of productivity gains or on the role of productivity analysis in the corporate budget, and it does not deal with productivity comparisons.) This objective is served by a brief discussion on methodological considerations (Section 2), an elaboration on some of the major measurement problems (Section 3) and a detailed account of how the component variables are measured in Bell Canada (appendices). The second objective is to describe and analyze Bell Canada's productivity performance during 1952 to 1979. This is accomplished in Section 4.

2. Methodology

At the expense of some simplification, the various approaches to productivity measurement that have been proposed in the literature can be classified in two broad categories; namely,

- the indexing approach, and
- the econometric approach[1] .

The indexing approach measures productivity gains as the difference between the aggregate growth rates of output and input and utilizes index number formulae to obtain the aggregates, while the econometric approach uses additional information about the structure of the production process, derived from the parameter estimates of production or cost functions. The indexing approach views productivity gains as a residual of output growth which cannot be explained by proportional growth in inputs. The econometric approach attempts to measure productivity gains by component.

Both approaches are based on the neoclassical theory of production. For a brief summary of the underlying theoretical issues,[2] let us begin with a general transformation function,

$$H(Q_{1t}, \ldots, Q_{nt}; X_{1t}, \ldots, X_{mt}; t) = 0 \quad ,$$

in which Q_{it} $(i=1,\ldots,n)$ and X_{jt} $(j=1,\ldots,m)$, respectively, refer to outputs and inputs at time t. Aggregates of output and input at time t can be denoted by Q_t and X_t and their proportionate rates of growth by $\dot{Q}$ and $\dot{X}$.[3] The productivity gain is then defined as

$$\dot{PR} = \dot{Q} - \dot{X} \quad .$$

The condition for measuring productivity gains in this manner is that the transformation function is of the homothetic weakly separable form; i.e., it can be written as

$$H(Q_{1t}, \ldots, Q_{nt}; X_{1t}, \ldots, X_{mt}; t)$$

$$= H'[G'(Q_{1t}, \ldots, Q_{nt}), F'(X_{1t}, \ldots, X_{mt}, t)]$$

$$= H''[G''(Q_t), F''(X_t, t)]$$

The traditional production function

$$G(Q_t) - F(X_t, t) = 0$$

is obtained, when the homothetic separability is additive.

Different index number formulae correspond to different forms of the production function F. The index number which is equal to Q_t/Q_{t-1}, derived from a specific F production function is called an exact index number for F. Diewert (1976) shows that the Törnqvist volume index (used in the Bell Canada productivity study) is exact for a homogeneous translog production function and the Törnqvist price index is exact for a homogeneous translog cost function. Furthermore, Diewert defines an index number as superlative if F for which it is exact can provide a second-order approximation to an arbitrary linearly homogeneous production function, and shows that the Törnqvist index is superlative, while the Laspeyres and Paasche indices (on which the Kendrick productivity index of the Bell Canada productivity study is based) are not.

Productivity gains are regarded as the consequence of technical progress and several authors[4] have established the conditions under which the residual or index number measure of productivity coincides with the effect on output (or cost) of technological changes. The conditions are (1) constant returns to scale and (2) perfect competition in the input and output markets.

Abramovitz (1956), Fabricant (1959), Kendrick (1961), Denison (1962) and others emphasized that residually measured productivity gains capture, in addition to technical progress, a multitude of systematic and random effects and that they are "a measure of our ignorance". There are two sources of other influences; viz., random distrubances in the output or cost of the firm and violations of the above mentioned assumptions. It was recognized that errors in optimization, economies of scale and measurement errors are parts of the residual productivity gain. Some monopoly phenomena (cross-subsidized prices and rate of return constraint) were added to the list of components in recent years.[5] There is reason to believe that especially strong violations of the assumptions of perfect competition and constant returns to scale exist in the case of a regulated public utility, like Bell Canada. Econometric cost studies indicate that Bell's technology is characterized by increasing returns to scale.[6] Over the study period, the product market has been largely monopolized and marginal cost has not been a "rate setting objective".

Attempts to decompose residually measured productivity gains by econometrically estimated components by Griliches, Denny, Fuss, Waverman, Everson, Nadiri, Schankerman[7] constitute what is called the econometric approach to productivity measurement. This approach estimates production or cost functions in which some of the restrictive assumptions of the indexing approach are relaxed, phenomena such as increasing returns to scale, non-marginal-cost pricing or rate of return regulation are modeled, and the parameter estimates of the function are used to measure components of productivity gains. E.g., productivity gains due to scale economies are measured with the aid of the scale or cost elasticity and the index of input or output. The sum of the components is the econometric measure of productivity gains. It excludes the effect of random disturbances in the output or cost and includes the systematic effects only.

Both approaches generate useful information. The all-inclusive nature of residually measured productivity gains makes the indexing approach valuable. The productivity indices show how the overall efficiency of production changes as the ultimate result of a great number of events, which influence the company's operations. Analysis of the impact of productivity improvements on costs, profits, factor usage, output prices, etc., utilize the residual measure of productivity gains. The econometric approach, on the other hand, aids the analysis and evaluation of productivity gains by identifying their causes and quantifying their components. The econometric approach attempts to determine how much productivity improvement is due to technological innovations, economies of scale and other factors. In doing so, it provides opportunities to distinguish between controllable and uncontrollable changes in efficiency and to evaluate the influence on productivity gains of policy, regulatory and managerial decisions involving a wide range of issues, such as industry structure, telephone rates, demand growth, innovations, etc.

2.1 The Indexing Approach

The Bell Canada productivity study uses two index number formulae. The simpler of the two is Kendrick's arithmetic productivity index,[8] which is intuitively appealing; thus, its use is quite widespread, even though the production model on which it is based appears to be too restrictive for Bell Canada.

The second formula is based on the discrete log-change Törnqvist-Theil approximation[9] to the time-continuous Divisia indices of output and input and is referred to as the Törnqvist index of productivity. Output and input prices are utilized in the process of aggregating individual outputs and inputs in this formula under the assumption that the revenue share of each output is equal to the cost elasticity with respect to the same output and that the cost share of each input is equal to the output elasticity with respect to the same input. The index has generally favourable properties. It is invariant, approximates the factor reversal test (very closely for Bell Canada) and satisfies all the other conventional index tests.

Among its limitations, three problems deserve attention. First, although the continuous Divisia index is reproductive (the Divisia index of Divisia indices is also a Divisia index), the Törnqvist-Theil approximation is not, unless all the aggregator functions are in linearly homogeneous translog form.[10] There is only a small difference between the one-step and two-step output and input aggregates for Bell Canada. All reported Törnqvist indices of productivity are based on one-step output aggregates and two-step input aggregates.

Secondly, the Bell Canada productivity study uses annual observations, hence a problem of approximation exists. The discrete index is asymptotic to the continuous index; i.e., the closer the observation points the better the approximation. Unfortunately, the high cost of collecting monthly data prevents us from solving the approximation problem.

The third problem is that of possible path dependence of the Divisia index. Hulten (1973) established three conditions for the path independence of the Divisia line integrals. The first condition, the existence of an aggregate, is normally assumed. However, in the case of Bell Canada, Smith and Corbo (1979) tested and rejected the hypothesis of weak separability of inputs from outputs in a multi-output multi-input translog cost function, suggesting that a single output aggregate cannot be formed.[11] The second condition, constant returns to scale, is generally not satisfied in econometric production and cost models of Bell Canada.[6,12] The third condition is no more than the mathematical implication of behaviour optimization.

In an attempt to preserve path independence in the case of an increasing-returns-to-scale technology , Nadiri and Schankerman (1979) introduced the quasi-Divisia index of input volumes recommended by Hulten (1973). Hulten's index number formula is not pursued in the Bell Canada productivity study.

Productivity in any year (PR_t) is defined as the ratio between output (Q_t) and input (X_t) in the same year; i.e., $PR_t = Q_t/X_t$; and the productivity index between a base year (B) and any year t is

$$(1) \quad \frac{PR_t}{PR_B} = \frac{Q_t/Q_B}{X_t/X_B} \quad .$$

Kendrick indices rely on Laspeyres volume indices of n outputs and m inputs, with the respective output prices (P) and input prices (W) as weights. The formula is

$$(2) \quad \left(\frac{PR_t}{PR_B}\right)_K = \frac{\sum_i (P_{iB}\ Q_{it})}{\sum_i (P_{iB}\ Q_{iB})} \Bigg/ \frac{\sum_j (W_{jB}\ X_{jt})}{\sum_j (W_{jB}\ X_{jB})} \quad (i=1,\ \ldots,\ n;\ j=1,\ \ldots,\ m),$$

and in the case of output exhaustion in the base year ($\Sigma(P_{iB}\ Q_{iB}) = \Sigma(W_{jB}\ X_{jB})$), it becomes

$$(3) \quad \left(\frac{PR_t}{PR_B}\right)_K = \frac{\sum_i (P_{iB}\ Q_{it})}{\sum_i (W_{jB}\ X_{jt})} \quad .$$

Until 1979, the Kendrick indices were calculated using 1967 as the base year. However, since labour price increased faster than the price of capital and the capital/labour ratio increased throughout the observation period, which begins in 1952, the 1967-based Kendrick index had a decreasing upward bias through time between 1952 and 1967 and an increasing downward bias through time from 1969 onward. To eliminate the bias, a moving B = t-1 base year was introduced in 1979. The annual indices are chained and the resulting series is normalized around its 1967 value to make it comparable with the 1967-based indices. The moving-base-year index is consistent with the Laspeyres volume index solution of Kendrick (1961).

The Törnqvist formulae of output and input volume indices are

$$(4)\quad \left(\frac{Q_t}{Q_{t-1}}\right)_T = \underset{i}{\pi} \left(\frac{Q_{it}}{Q_{i,t-1}}\right)^{\frac{1}{2}(r_{it} + r_{i,t-1})}$$

$$(5)\quad \left(\frac{X_t}{X_{t-1}}\right)_T = \underset{j}{\pi} \left(\frac{X_{jt}}{X_{j,t-1}}\right)^{\frac{1}{2}(s_{jt} + s_{j,t-1})}$$

where r and s refer to revenue and cost shares, respectively. E.g., $r_{it} = (P_{it}\ Q_{it}) \ / \ \sum_i (P_{it}\ Q_{it})$ and $s_{jt} = (W_{jt}\ X_{jt}) \ / \ \sum_j (W_{jt}\ X_{jt})$.

The Törnqvist index of productivity is

$$(6)\quad \left(\frac{PR_t}{PR_{t-1}}\right)_T = \left(\frac{Q_t}{Q_{t-1}}\right)_T \Big/ \left(\frac{X_t}{X_{t-1}}\right)_T .$$

Hulten's index of productivity differs from the Törnqvist index in that its input weights are $\frac{1}{2}(z_{jt} + z_{j,t-1})$, where $z_{jt} = (W_{jt}\ X_{jt}) / \sum_i (P_{it}\ Q_{it})$ and $z_{j,t-1}$ is similarly defined.

2.2 The Econometric Approach

Following Denny, Fuss, Everson (1979) and Denny, Fuss, Waverman (1979), an econometric estimation of productivity gains through a technology effect and a scale effect is described below. Although the estimation is simplified in that it uses a single output model and does not quantify effects such as non-marginal-cost pricing, it has produced more reasonable empirical results than more elaborate productivity compositions.

The neoclassical production function is

$$(7)\quad Q = f(X_1, \ldots, X_m, t),$$

where Q, X and t denote output, input and technological changes, respectively. It yields a definition of output shifts caused by technological changes. The proportional output shift is

$$\dot{A} = \frac{\partial \log Q}{\partial t} .$$

The "econometric" productivity gain can be defined as the sum of the technology effect and the scale effect; i.e.,

$$(8)\quad (\dot{PR}) = \dot{A} + (\varepsilon - 1)\,\dot{X} ,$$

where ε is the scale elasticity, estimated in the production function as

$$\varepsilon = \sum_j \frac{\partial \log Q}{\partial \log X_j} \quad (j=1, \ldots, m) ,$$

and $\dot{X}$ denotes the Divisia index of aggregate input, i.e.,

$$\dot{X} = \sum_j \frac{W_j\, X_j}{\sum_j (W_j\, X_j)}\, \dot{X}_j ,$$

where W denotes input prices.

In the dual cost function of the form

$$(9)\quad C = g(W_1, \ldots, W_m, Q, t) ,$$

the cost shifts that are caused by technological changes are defined as

$$\dot{B} = \frac{\partial \log C}{\partial t}$$

and the productivity gains are expressed again as the sum of the technology effect and the scale effect; i.e.,

$$(10)\quad (\dot{PR}) = -\dot{B} - (\varepsilon_{CQ} - 1)\,\dot{Q} ,$$

where ε_{CQ} is the cost elasticity with respect to output, estimated in the cost function as

$$\varepsilon_{CQ} = \frac{\partial \log C}{\partial \log Q} ,$$

and $\dot{Q}$ is the Divisia index of aggregate output; i.e.,

$$\dot{Q} = \sum_i \frac{P_i \, Q_i}{\sum_i (P_i \, Q_i)} \dot{Q}_i \quad .$$

Equations (8) and (10) yield alternative econometric measures of productivity gains. In equation (8), a production function provides the estimates of $\dot{A}$ and the scale elasticity, ε, and the input index is used. In equation (10), the estimates of $\dot{B}$ and the cost elasticity, ε_{CQ}, are derived from a cost function, and the output index is utilized. Since the shifts in the production function and the cost function are related through the cost elasticity, $-\dot{B} = \varepsilon_{CQ} \dot{A}$,[13] and the scale and cost elasticities are also related, $\varepsilon = \varepsilon_{CQ}^{-1}$, it is possible to use an estimated cost function ($\dot{B}$ and ε_{CQ}) both with the input index, as in the equation

$$(11) \quad (\dot{PR}) = - \dot{B}/\varepsilon_{CQ} + (\varepsilon_{CQ}^{-1} - 1) \dot{X} \quad ,$$

and with the output index, as in equation (10), to arrive at econometric estimates of productivity gains.[14]

The productivity gains of equations (10) and (11) are equal if the actual input growth of the company is equal to the growth of the cost minimizing total input. However, favourable or unfavourable production conditions and errors in cost minimization can make the actual input growth either slower or faster than the cost minimizing input growth.

When the estimated cost function does not have time as a variable, but contains a proxy (T) for technological changes, the intertemporal cost shifts can be obtained as

$$(12) \quad \dot{B} = \frac{\partial \log C}{\partial \log T} \cdot \frac{\partial \log T}{\partial t} \quad .$$

Since the Törnqvist volume indices of output and input that yield the residually measured productivity gains ($\dot{PR}$) represent a discrete approximation to the continuous Divisia indices, the continuous formulae on the right hand side of equations (10) and (11) should be approximated as well. A linear approximation of the technology effect and the scale effect can be obtained with ease from an estimated translog cost function. Taking equation (10) as an illustration,[15] the technology effect ($-\dot{B}$) is approximated by component, as they are shown in (12), in the formula

$$(13) \quad -\dot{B}_t = -\tfrac{1}{2} (\varepsilon_{CT,t} + \varepsilon_{CT,t-1}) (\log T_t - \log T_{t-1})$$

where $\varepsilon_{CT} = \partial \log C/\partial \log T$ is the technology elasticity of cost.

As $\dot{Q} = \partial \log Q/\partial t$ becomes $\dot{Q}_t = \log Q_t - \log Q_{t-1}$, the scale effect is expressed in the discrete case as

$$(14) \quad -\dot{E}_t = - [\tfrac{1}{2} (\varepsilon_{CQ,t} + \varepsilon_{CQ,t-1}) - 1] (\log Q_t - \log Q_{t-1}) \quad .$$

When the underlying cost function is a generalized translog (GTL) in which the Box-Cox transformation is applied to the output and technology proxy variables, the approximation becomes more complicated. I am indebted to Professor Melvyn Fuss for the following solution.

The Box-Cox transformation of the technology proxy variable results in $T^* = (T^{\lambda_T} - 1)/\lambda_T$ and the cost shift caused by T^*, $\Gamma = \partial \log C/\partial T^*$, is directly estimated. Since Γ is a component of the technology elasticity of cost, (12) can be expressed in a GTL-specific form as

$$(15) \quad \dot{B} = \frac{\partial \log C}{\partial T^*} \cdot \frac{\partial T^*}{\partial \log T} \cdot \frac{\partial \log T}{\partial t} \quad .$$

Linear approximation can be applied to Γ and the other terms on the right hand side of (15) are readily approximated by

$$\Delta T^* = (T_t^{\lambda_T}-1)/\lambda_T - (T_{t-1}^{\lambda_T}-1)/\lambda_T = (T_t^{\lambda_T} - T_{t-1}^{\lambda_T})/\lambda_T \quad ,$$

$$\Delta \log T = \log T_t - \log T_{t-1} \text{ and}$$

$$\Delta t = 1 \quad .$$

Hence the generalized translog cost function yields the following discrete approximation to the technology effect in (10):

$$(16) \quad -\dot{B}_t = -\tfrac{1}{2}(\Gamma_t + \Gamma_{t-1}) \frac{T_t^{\lambda_T} - T_{t-1}^{\lambda_T}}{\lambda_T} \quad .$$

The scale effect is treated in a similar manner. As the Box-Cox transformation of the output variable generates $Q^* = (Q^{\lambda_Q}-1)/\lambda_Q$, the cost shift caused by Q^*, $\Phi = \partial \log C/\partial Q^*$, is directly estimated and it is a component of the output elasticity of cost, the scale effect in (10) can be written in a GTL-specific form as

$$(17) \quad -\dot{E} = -\left(\frac{\partial \log C}{\partial Q^*} \cdot \frac{\partial Q^*}{\partial \log Q} - 1 \right)\dot{Q} \quad .$$

The discrete approximation of (17) is

$$(18) \quad -\dot{E}_t = -\left[\tfrac{1}{2} (\Phi_t + \Phi_{t-1})\frac{(Q_t^{\lambda_Q} - Q_{t-1}^{\lambda_Q})/\lambda_Q}{\log Q_t - \log Q_{t-1}} - 1\right](\log Q_t - \log Q_{t-1}) \quad .$$

Productivity gains resulting from the indexing approach contain (1) the effect of scale economies, (2) the effect of technological changes and (3) other effects. Because they exclude other effects, the econometric measures as defined in equations (10) and (11) do not fulfill the role of a single-number measure of changes in overall productive efficiency; thus, they do not constitute an alternative to, but rather a component of, residually measured productivity gains.

With the indices of output, input and productivity given and the cost elasticity derived from an estimated cost function, there are several ways to arrive at the composition of productivity gains. When the technology effect is obtained as the residual of Törnqvist gains after subtracting the scale effect (with either $\dot{Q}$ or $\dot{X}$), the other effects, which are related to the residual term of the estimated cost function, are attributed to technology. They can bias the annual estimates of technology effect either upward or downward but, because they have an approximately zero mean, the long-run (sample period length) average technology effect remains unbiased. Another problem is that the error in the cost elasticity estimate influences not only the scale effect in which it appears, but also the residual technology effect. When the decomposition is done at the bounds of the 95% confidence interval of ε_{CQ}, the results can change rather drastically and the technology effect may become negative in several years at the lower bound of the cost elasticity.[16]

The solution pursued in this paper defines the technology and scale effects as in equation (10) and separates out other effects into a residual (R_t) of Törnqvist productivity gains;[17] thus, the composition becomes

$$(19) \quad (\dot{PR})_t = -\dot{B}_t - \dot{E}_t + R_t \quad .$$

The underlying cost function is specified as a single-output generalized translog with a logarithmic output variable and Box-Cox transformation of the technology proxy. Hence $-\dot{B}_t$ is defined as in (16) and $-\dot{E}_t$ is approximated according to (14).

3. Some Measurement Problems

3.1 Output Measurement

Aggregate output volumes are represented by deflated or constant dollar revenues in the Bell Canada productivity study. Deflated revenues encounter the usual problems associated with the appropriateness of prices as weights (cross-subsidized and subsidizing prices may distort the output aggregates) and, in the case of Bell Canada, flat monthly rates for local services present a special problem in that they reduce the sensitivity of the measure to output changes, because changes in usage are not reflected in deflated revenues.

Deflated revenues also present a number of advantages in measuring output volumes. Perhaps the most important advantage is that the availability of prices for individual services makes a detailed and elaborate aggregation procedure possible, while the theoretically more suitable cost weights are not available and are difficult to approximate, even for large service aggregates. Another desirable aspect of deflated revenues is that they are capable of reflecting changes in the quality of services, when these are accompanied by telephone rate changes.[18]

The ideal output measure would express service volumes in physical units. For local services, it could be based on usage and consider the number, the duration and perhaps the distance of local calls. Qualitative adjustments for changes in access characteristics such as the size of the local calling area, the number of telephones in households and access-related terminal equipment features (dial, touchtone, preprogrammed) as well as for changes in other terminal equipment features should play an important role in local output volume measures, since the quality of service changes constantly and significantly. Patterns of the utility of local calls could probably be approximated reasonably well by the distribution of calls. Although there is no reason to declare the task of measurement in physical units impossible, no practically useful solution exists, nor is one expected in the foreseeable future.

The measurement of intra-Bell toll message volumes is more straightforward. Sufficient detail on the number, duration, distance, type, day and hour of calls is given and the aggregation is done with the only available weights - prices. The appropriateness of prices as weights seems to be the only problem with respect to intra-Bell calls. However,for inter-company toll messages, deflated revenues present a major difficulty, due to the fact that the price indices, calculated from individual Bell Canada rates, are weighted in most cases by Bell-originated call volumes, while Bell's output volumes and prices are thought to be better represented by the so-called settled revenues. Since the revenue settlement procedure does not result in price and volume information, it is not possible to calculate price indices with which settled revenues could be satisfactorily deflated.

With certain exceptions,[19] originated revenues are billed and collected by Bell Canada for long distance telephone messages originating within and terminating outside Bell Canada territory. Settled revenues differ from originated revenues because payments are made on the one hand by Bell Canada to other telephone companies to compensate them for the Bell-originated messages that terminate in or go through their network and on the other hand to Bell Canada by other telephone companies to compensate for carrying calls originated elsewhere. There are settlements with small independent telephone companies which represent only a small adjustment to originated revenues. These settlements are made bilaterally between Bell Canada and each of more than 50 independent companies. Other settlements are handled by the TCTS, with the exception of adjacent member settlements, which are bilateral. The total TCTS inter-company message toll revenue is distributed among member companies, after settlements with Telesat, Teleglobe, AT&T, etc., have been made. Settled revenues from adjacent member settlements are separated from TCTS settled revenues. However, the latter is not available separately for Trans-Canada, US and overseas.

Since originated revenues are available for each settlement, the total TCTS settled revenue is split to Trans-Canada, US and overseas by originated revenue ratios and the adjacent member settled revenues are added to the Trans-Canada settled revenues for the purposes of the productivity study.

Other toll services represent a mixture of the problems associated with local and message toll services. Some rates are based on usage and others are flat monthly rates. Occasionally, like in the case of TWX, the actual bill is a combination of usage sensitive and flat monthly charges. Revenue settlement exists for other toll services as well as for message toll. A distinct characteristic of other toll services is that equipment and facility rental constitutes a relative large part of revenues.

Another issue is the number of individual output categories. Ideally, prices and physical volumes for each individual service should be known and the aggregation should be done by consistently applying the chosen index number formulae. However, because of the enormity of the task of obtaining 30 to 40 thousand prices and volumes, the number of outputs has been reduced to the ten categories shown in Section 1.1 of Appendix A.

In order to maintain the consistency of the chosen index number formula in the process of output aggregation, Paasche and Törnqvist price indices and Laspeyres and Törnqvist volume indices should be available for each of the ten output categories. However, only Paasche price indices and Laspeyres volume indices are available at the present. As a result, the $Q_{it}/Q_{i,t-1}$ coefficients in the Törnqvist volume index formula (see equation (4)) are represented by ratios of deflated revenues, where the deflators are Paasche price indices.

A further issue is the definition of the contents of the output measure. The output of a firm is generally measured as the aggregate of the volumes of all products being produced. This all-inclusive aggregate is referred to as (real) gross production. If certain input separability conditions are fulfilled in the production function of the firm (i.e., when intermediate inputs are weakly separable from labour and capital), the intermediate inputs are removed from the output and input variables of the productivity measure. The resulting output variable (gross production, less intermediate inputs) is called (real) value added. The available empirical evidence suggests that the intermediate inputs of Bell Canada are not separable.[20] Although the productivity study maintains some value added measures, the following analysis of productivity gains is based entirely on real gross production.

The last item to be discussed in this sub-section is the treatment of revenue-related non-income taxes (RROT). Price indices (P_{it}) and constant dollar revenues, representing output volumes (Q_{it}) can be measured directly on the basis of unadjusted operating revenues. If this happens RROT either appears as part of the price of capital (when the residual rate of return is the measure) or it is left unaccounted for (if the user cost is applied).

Alternative measures are obtained when either output prices or output volumes are adjusted for the existence of revenue-related non-income taxes. In the output price adjustment, these taxes are regarded as direct deductions from the prices of telephone services. An adjusted Paasche price index of output can be given as

$$\hat{I}_p = \frac{\Sigma(\hat{P}_{it}\ Q_{it})}{\Sigma(\hat{P}_{i,t-1}\ Q_{it})} ,$$

where $\hat{P}_{it} = P_{it}\ (1 - \frac{RROT_t}{\Sigma(P_{it}\ Q_{it})})$; i.e., where the original price (P_{it}) is lowered by the value of total revenue-related non-income tax ($RROT_t$), prorated among output items according to their revenue shares, per unit of output volume. The $\hat{P}_{it}$ formula implies that the relationship between the unadjusted and adjusted Paasche price indices is

$$\hat{I}_p = I_p \frac{1 - \rho_t}{1 - \rho_{t-1}}$$

where I_p is the unadjusted index and ρ is the rate of revenue-related non-income tax; i.e., $\rho_t = RROT_t/\Sigma(P_{it}\ Q_{it})$.

Under the assumption that municipal inputs provided for Bell Canada at the expense of the revenue-related non-income taxes are intermediate inputs, an alternative adjustment of constant dollar revenues by output category can be made. $RROT_t$ is deflated by the Paasche price index of total output and the constant dollar tax is prorated among output categories according to their constant dollar revenue shares. The resulting adjusted constant dollar revenue for output category i is

$$P_{iB}\ \hat{Q}_{it} = P_{iB}\ Q_{it}\ (1 - \rho_t).$$

Since revenue-related non-income taxes represent only a very insignificant output price or volume adjustment, the choice of their treatment in the productivity study has no empirical importance.

A comparison of the Laspeyres and Törnqvist volume indices, reported in Tables B.2 and B.3 (Appendix B), reveals that little is gained in Bell Canada's productivity measurement by using Törnqvist volume indices for output. The average annual rate of growth in gross production is 8.67% (Laspeyres) or 8.77% (Törnqvist) for the period 1952 to 1979. The negligible understatement of output growth by Laspeyres indices results from a very slight overstatement of toll growth and understatement of local growth. The differences between the two indices might well be larger if they were built from individual prices and volumes. However, since only a single set of category level price indices is available, the difference between the Laspeyres and Törnqvist aggregations within each of the 10 categories cannot be captured.

The output volume index tables illustrate some well known facts: e.g., that local output volumes have grown more slowly than toll, and message toll has grown more slowly than other toll. The growth of local service volumes is characterized by a slowdown, as very high growth rates existed during the 1953 to 1959 period and growth was very slow in recent years. No trend is observable for the period 1960 to 1975. Growth after 1975 was slower than in any other sub-period. There are some rather drastic year-to-year fluctuations in the growth rates of monopoly toll services, with no underlying strong trend. Output grew slowly between 1957 and 1961 and fast growth is observable during the 1972 to 1975 period. Very high growth rates of competitive toll services were achieved in private line on small volumes at the beginning of the observation period. This is a typical new product phenomenon. Later, new services such as TWX and some data services boosted up the growth rates several times. It is interesting to observe that competitive toll grew more slowly than monopoly toll after 1970.

The growth pattern of gross production is dominated by year-to-year fluctuations and the sample period cannot be broken down into analytically useful sub-periods. Two periods of below-average growth are 1958 to 1961 and 1976 to 1979, while the only longer period with above-average growth lasted from 1972 to 1975.

3.2 Labour Measurement

Labour input is measured by the number of hours worked directly on the production of telecommunication services. Several issues deserve attention. First, the Bell Canada labour force is classified according to occupational groups and length of service. Denison (1962) and Gollop and Jorgenson (1980) included labour classes distinguished according to age, sex and education in their classification of the US labour force. Gollop and Jorgenson suggest that classification according to various demographic characteristics is desirable and Denison argues that different personal characteristics also should be considered. Since the Bell Canada productivity study does not reflect labour characteristics other than occupation and experience in the labour input, these characteristics are captured in the measured productivity changes.

Secondly, in the absence of a suitable measure, hours worked by management and clerical employees on the production of regulatory and other information, not directly related to the production of telecommunications services, are included in the labour input measure. The output of these hours is not accounted for in the output measures of Bell Canada; thus, the inclusion of information-producing hours lowers/increases productivity gains, when information-producing hours grow faster/slower than other inputs.

Thirdly, a certain percentage of Bell Canada's labour force is employed in the process of constructing telephone plant rather than directly producing telephone services or managing the company. Since the value of their labour input is included in the value of the resulting telephone plant, the hours they work are excluded from the labour input measure. However, the percentages of expensed and capitalized labour that are derived from company records have been altered on several occasions by changes in accounting procedures and these changes have influenced the measured productivity gains.

Finally, it is assumed that labour is compensated in proportion to its marginal revenue product and the qualitative differences that cause marginal products to vary are accounted for in labour volume indices, using hourly labour cost by category. Since the marginal products are not measured and, at least at the present time, cannot be satisfactorily estimated, it is not known how much distortion the hourly labour cost weights cause in the volume indices of labour input.

Labour-related non-income taxes (LROT) are applied as an upward adjustment of labour price. The adjusted labour price is

$$\hat{W}_{it} = W_{it}(1 + \rho_t^L) \quad ,$$

where ρ^L is the rate of labour-related non-income taxes; $\rho^L = \mathrm{LROT}/\Sigma(W_i L_i)$. The same kind of adjustment is made to the price of capital input.

Volume and price indices for Bell Canada's labour input are shown in Tables B.6 and B.7 (Appendix B), respectively. A comparison of Laspeyres and Törnqvist indices suggests that the choice of the index number formula generally does not affect the measured changes in labour input to a significant degree.

During the sample period, total expensed hours worked grew only by 1.5% per year. This slow growth is due mainly to the substantial reduction in the number of telephone operators. The Laspeyres index with constant 1967 weights shows a 2% growth per annum. The variable-weight Törnqvist index estimates the growth rate for total labour input lower at 1.9% p.a., with an implied increase in input per hour of .4% p.a. There are five clearly distinguishable sub-periods, as shown in the table below.

TABLE 1: Average Annual Growth Rates of Aggregate Labour Input

PERIOD	UNADJUSTED HOURS WORKED	ADJUSTED HOURS WORKED		LABOUR INPUT PER HOUR	
		LASPEYRES	TÖRNQVIST	LASPEYRES	TÖRNQVIST
1952-57	5.27%	5.17	5.54	-.10	.27
1957-62	-3.79	-1.99	-3.21	1.80	.58
1962-66	3.11	2.39	3.06	-.72	-.05
1966-72	-.93	.02	-.10	.95	.83
1972-79	4.11	4.35	4.34	.24	.23
1952-79	1.54	2.04	1.94	.50	.40

The first period (1952 to 1957) is characterized by sizable increases in expensed hours. The Laspeyres index shows a 5.2% and the Törnqvist index shows a 5.5% average annual growth in labour input. It is interesting to observe that the Laspeyres index implies a small decline and the Törnqvist index implies a moderate increase in input per hour. In this 5-year period, the fastest growing occupational groups were clerical and other non-management as well as other management. Hours worked by telephone operators on the other hand grew slowly at a rate of 1.6% per year. Hours worked by foremen and supervisors declined slightly.

Between 1957 and 1962, the number of expensed hours worked declined by 3.8% p.a. The labour input decline is 2% in the Laspeyres formula and the Törnqvist index shows a 3.2% average annual decline. The quality-generated increase in input per hour is 1.8% p.a. according to the Laspeyres index and only .6% in the Törnqvist index. This is the fastest quality mix improvement in the entire sample period. It was caused by a very fast decline in operator hours and a moderate increase in other management.[23] Hours worked in all full-time groups also declined, while a small increase is observed in the part-time employee group. The substantial reduction in operator hours coincides with, and is largely caused by, the introduction of Direct Distance Dialing (DDD).

In the 1962 to 1966 period, expensed hours worked grew again, at a rate of 3.1% per year. Both indices indicate that the quality mix of the labour force of Bell Canada shifted toward lower quality labour and input per hour declined as a result. The decline occurred despite a slow growth of hours worked by telephone operators and fast growth in other management hours.

During the years 1966 to 1972, a decline in the number of hours worked was accompanied by strong growth in input per hour. Operator hours declined substantially (by 4.3% p.a.) and other management hours continued to increase, though at a lower rate than in the preceding period.

The last seven years (1972 to 1979) produced a 4.1% average annual increase in expensed hours worked. This rate is higher than at any time after 1957 but lower than the rates that prevailed before 1957. Operator hours declined in this period but hours worked in all other full-time occupational groups increased substantially.

3.3 Capital Measurement

The capital input volume and price measures are conceptually analogous to the volume and price of labour input.[21] Capital input is represented by the constant dollar stock of capital. The measure is often referred to as reproduction cost, signifying that technological changes in equipment manufacturing, resulting in costless quality improvements, are not allowed to lower the price of capital. For more on this subject, see Denison (1957) and Usher (1980). Capital input price is measured by either the residual rate of return or the user cost of capital. Section 3.2 of Appendix A elaborates on the capital price measures.

Capital stock is used instead of utilized capital out of necessity rather than due to theoretical considerations. Suitable capital utilization measures are not available and the adjustment procedures that are recommended in the literature, e.g., Griliches and Jorgenson (1966), Berndt and Wood (1977) or Gollop and Jorgenson (1980), are not applicable for Bell Canada. However, utilization adjustment is a debatable issue. Many pros and cons have been discussed both within and outside the productivity measurement debate between Denison and Griliches and Jorgenson. E.g., Kendrick (1973) states that "The degree of capital utilization reflects the degree of efficiency of enterprises ... Hence, in converting capital stocks into inputs, we do not adjust capital for changes in rates of capacity utilization, and thus these are reflected in changes in the productivity ratios".[22]

If the total annual capital cost (C_K) is observed (e.g., the residual return to capital) the measure of capital price is affected by the use of capital stock (K). The price of capital stock is $k = C_K/K$, while the price of utilized capital (K') is $k' = C_K/K'$, where $K'<K$, hence $k'>k$.

Capital stock can be defined narrowly as telephone plant in service or more broadly by including plant under construction or even the so-called working capital. The inclusion of plant under construction reduces the reliance of the production model on the technologically determined input-output relationship, since PUC is related to a future flow of outputs rather than to contemporaneous output volumes. In other words, the given technology is better reflected if PUC is excluded from capital. However, it is managerially meaningful to reflect changes in PUC in the measured productivity gains. As an alternative, the productivity measure so derived allows management to see how the relationship between outputs produced and all resources used up changes through time, regardless of how and why the resources were consumed. Similar reasons lead to the inclusion of working capital in the capital input measure. The resulting alternative measures reflect the impact of financial items (cash, short-term deposits, accounts receivable and accounts payable) and inventory changes on productivity gains.

Since the volumes of plant under construction and working capital are very small in comparison with telephone plant in service, the difference between the narrow and the broad definitions of capital does not result in a significant alteration of the empirical conclusions in the Bell Canada productivity study.

The main problem with estimating the economic depreciation of Bell's capital is that the fall in the market value of telecommunication equipment over time, due to simple aging, physical deterioration and obsolescence, is not observable, because there is very limited market for used equipment. However, as a surrogate, accounting depreciation rates are used to represent the degree of economic depreciation.

Table B.10 contains 1967-based indices of unadjusted net capital stock as well as three alternative Törnqvist volume indices. The figures suggest three conclusions. First, PUC is too small, compared to the volume of plant in service, to alter the growth pattern perceivably. Second, the growth patterns of the four series are almost identical. Third, the growth of capital has been slowing down during the period of observation.

The almost identical growth patterns of adjusted and unadjusted capital suggest that there was no quality mix change during the period of observation. This result requires further analysis and Tables 2 and 3 have been assembled to aid this analysis. The figures in Table 2 reveal a very substantial mix change in net plant volumes. The share of central office equipment (COE) increased rather sharply from 22.4% in 1952 to 35.4% in 1979. The share of every other plant category declined, though the decline was negligible in land and general equipment. The most pronounced decline took place in the share of outside plant. Table 3 shows user costs for major plant classes. The arithmetic mean of class values uses weights taken from Table 2. COE has a quality indicator [23] which is close to the average; thus, changes in its share in total net plant do not have a significant effect on average quality. Land and general equipment are too small and their effects offset each other, so that they do not have a significant quality effect either. Below-average quality buildings and outside plant had declining shares, which resulted in quality mix improvement. Above-average quality station equipment also had declining shares and its impact is to lower average quality. The offsetting effects of station equipment on the one hand and of buildings and outside plant on the other hand explain the almost identical growth patterns of adjusted and unadjusted net capital. Laspeyres-indexing of capital volumes has not been attempted in the Bell Canada productivity study.

Chart 1 below depicts the annual growth rates of total net capital (excluding PUC) and those of the net value in constant (1967) dollars of central office equipment. The growth of net capital stock has been slowing down during the 1952 to 1979 period. The annual growth rates have a pronouncedly linear downward trend with fairly small annual variation and two bulges. The first bulge appears in the period 1955 to 1962 and the second one during the years 1974 to 1978. The first bulge is associated with the heavy investments necessitated by DDD and the second one, while it may be more complex in nature, appears to be associated mainly with the rapid shift to electronic equipment. Both bulges seem to be related largely to significant changes in switching technology. The pattern of COE volume growth supports this conclusion as it approximates closely (albeit with greater variation and a local minimum in 1966) the pattern of total net capital growth.

TABLE 2: Net Capital Mix in 1952 and 1979

CONSTANT DOLLAR NET PLANT	PERCENTAGES	
	1952	1979
Land	1.48	1.15
Buildings	11.48	8.00
Central Office Equipment	22.43	35.41
Station Equipment	18.28	16.70
Outside Plant	42.76	35.53
General Equipment	3.57	3.21
TOTAL	100.00	100.00

TABLE 3: User Costs in 1952 and 1979

PLANT CLASS	USER COST	
	1952	1979
Land	4.59%	23.03%
Buildings	5.15	25.97
Central Office Equipment	10.70	33.01
Station Equipment	11.21	46.20
Outside Plant	6.75	28.04
General Equipment	10.48	45.21
Mean	8.37%	33.16%

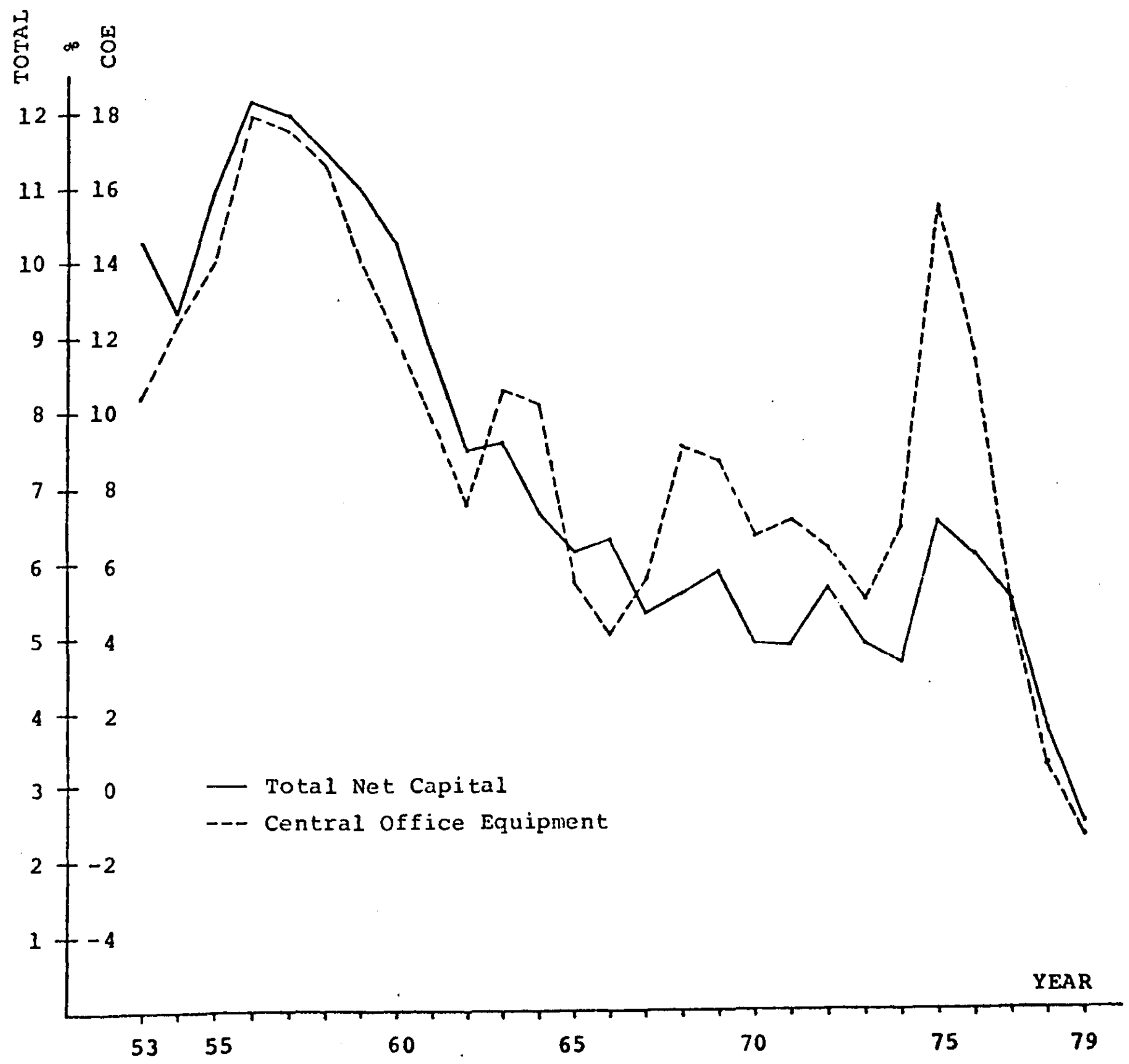


CHART 1: Annual Growth Rates of the Net Stock of Physical Capital and Constant (1967) Dollar Net Central Office Equipment

3.4 Material Measurement

Material input is measured by the constant dollar value of a great number of miscellaneous inputs such as materials, rents, supplies and services. Laspeyres and Törnqvist volume indices and implicit Paasche and Törnqvist price indices are calculated.

The non-income tax adjustment of materials differs from the method of upward adjustment of labour and capital prices. Since the amount of material-related non-income tax is either zero (1952-1966) or very small (1967-1979), it is simply given an assumed price index and treated as another material category.

The 1967-based price and volume index series are given in Table B.12. As explained in Appendix A, the GNE deflator was used to deflate the current dollar material, etc. cost series for both the Laspeyres and Törnqvist measures; thus, the two index series are identical during the period 1952 to 1969. It is interesting to observe that only very small deviations exist between the Laspeyres and Törnqvist volume index series during the period 1970 to 1979, indicating very little empirical gain from the Törnqvist formula.

4. The Productivity Performance of Bell Canada

Table 4 contains the information required for the following description and analysis of the productivity gains of Bell Canada during the period 1952 to 1979. The first two columns show annual productivity gains generated by technological changes and economies of scale, respectively. The third column presents residual productivity gains, due to factors other than technological changes and scale economies. The 1979 figures in the first three columns are preliminary.[24] The fourth column of the table demonstrates the actual productivity gains of Bell Canada. Törnqvist output and input volume indices were used to obtain a measure of actual gains.[25] The output index is that of real gross production, the capital measure is narrowly defined (plant in service) and the user cost of capital is utilized in the process of capital aggregation. Table 4 reflects in its structure the components of equation (19). The alternative decomposition formula yields somewhat different numerical results, but the differences do not alter any of the following conclusions.

The cost function from which the estimates of technology shift and output elasticity of cost were obtained is a single-output homothetic generalized translog cost function with logarithmic output and Box-Cox technology variable transformations. It appears in Kiss, Karabadjian and Lefebvre, 1981 (pp. 24-25 and Appendix B).

The actual average annual productivity gain of Bell Canada was 3.5% during the entire 27-year period of observation. The annual productivity gains appear to be generally very high.[26] Several sub-periods are distinguished by the pattern of annual gains. Table 5 gives a summary of period-average productivity gains.

The productivity gain of 1953 was slightly below the long-term average. Technological changes and scale elasticities resulted only in very small productivity improvement, but other circumstances were favourable, as indicated by the positive residual term in Table 4.

TABLE 4: Annual Productivity Gains and Their Composition

Year	Technology Effect	Scale Effect	Residual Effect	Actual Productivity Gains
1953	.4	.5	2.1	3.0
4	.6	.6	-.5	.7
5	-.1	.7	.2	.8
6	1.9	.9	-2.6	.2
7	2.3	1.1	.6	4.0
8	3.3	1.2	-2.4	2.1
9	1.5	1.7	2.1	5.3
1960	1.8	1.8	-.2	3.4
1	.9	2.2	1.7	4.8
2	.7	3.3	1.5	5.5
3	1.4	2.3	-2.9	.8
4	.9	2.9	-.5	3.3
5	.2	4.0	-.8	3.4
6	.3	4.2	-.5	4.0
7	.2	3.9	3.3	7.4
8	.7	3.5	.7	4.9
9	.5	4.6	-1.7	3.4
1970	.5	3.4	.5	4.4
1	.3	2.4	-2.6	.1
2	.4	4.5	1.7	6.6
3	.4	4.6	.2	5.2
4	.6	4.7	.2	5.5
5	.6	4.7	2.4	7.7
6	.4	3.3	-1.8	1.9
7	.4	2.9	-2.5	.8
8	1.4	3.6	-2.7	2.3
1979	1.3*	2.5*	-.4*	3.4

*Preliminary.

Very low productivity gains were registered in the following three years. Despite large increases in the size of the company's operations, the scale effect was very small, because only negligible scale economies existed in this period. Technological changes did not begin to contribute significantly to productivity gains until 1956. In fact, the technology change indicator declined slightly in 1955. The residual productivity gains indicate that the conditions were generally unfavourable for productivity improvement.

TABLE 5: Average Annual Productivity Gains

Period	Percentage Gain
1953	3.02
1954-56	.57
1957-71	3.77
1972-75	6.28
1976-79	2.09
1953-79	3.50

The revolution in switching technology, which started in 1956 with the introduction of the first crossbar central offices and customer-dialed long distance telephone calls (DDD), resulted in a suddenly very high direct technology effect on productivity gains. Technological changes also aided productivity improvement in indirect ways by increasing the degree of economies of scale and generating an upsurge in demand for long distance telephone services. As a result, the effect of economies of scale began to increase and reached very high levels by 1962-63. During the four years between 1963 and 1966, the average annual productivity gain generated by scale economies was approximately 3.3%. As the effect of the switching revolution gradually subsided and other circumstances were highly unfavourable (negative residual gain in each year), this 3.3% gain proved to be greater than the actual productivity gains of Bell Canada. The following years witnessed a continuation of small contributions to productivity gains by technological improvements, but the high degree of economies of scale kept productivity gains high. The residual gains show that favourable and unfavourable years alternated between 1966 and 1971. The entire 1957 to 1971 period is characterized by high rates of productivity improvement and fluctuations in the annual gains. The productivity gain was exceptionally high in 1967 (largely because of the high residual effect) and almost nonexistent in 1971. The poor productivity performance of 1971 was due mainly to an exceptionally large increase in material input and to a break in the output series, reflecting the establishment of Tele-Direct as a Bell Canada subsidiary.

The highest productivity gains of the period of observation were achieved between 1972 and 1975. Technological improvements contributed to productivity gains only very modestly, but very fast growth in demand for telephone, especially toll, services allowed the existing high degree of economies of scale to generate high productivity gains. Table 4 shows that the average annual productivity gain due to scale economies was 4.6% between 1972 and 1975. The residual gains were positive in each year and contributed significantly in 1972 and 1975.

The 1976-79 period represents a good example of the demand sensitivity of productivity gains. The actual gains were below the long-term average in all years and the average annual gain slipped to 2.1%. There are two major contributors to the erosion of productivity gains. First, demand for local and other toll services grew more slowly than in any other 4-year period during the observed 27 years and message toll demand also slowed down significantly. As a result, the contribution of scale to productivity gains dropped from 4.6% (1972-75) to approximately 3% per annum. Secondly, the slowdown in demand for telephone services coincided with, and was in part caused by, worldwide economic problems and some political uncertainties in Canada. Intensifying regulatory activities and some changes in accounting methods may also have had a significant negative impact on Bell Canada's productivity gains. The residual productivity gains have rather large negative values in this period.

A comparison of the actual and "econometric" productivity gains of Bell Canada in Chart 2 shows that the econometric measure captured the level and the essential features of the pattern of the company's productivity improvement, but left a substantial part of the annual variation of productivity gains unexplained. The unexplained variation is still helpful in the analysis of productivity gains to the extent that it identifies "favourable" years (e.g. 1967) and periods (e.g. 1972-75) and "unfavourable" years (e.g. 1971), and periods (1963-66 and 1976-79) for productivity improvements in Bell Canada.

Turning to the composition of the explained ("econometric") portion of productivity gains, Chart 3 shows that the contribution of scale economies was generally much greater than that of technological changes and that a certain pattern of relative contributions prevailed. High contributions were registered from technological changes (47 to 73% in 1956 to 1960) as a result of the introduction of crossbar central offices and DDD. As the new technologies gradually became dominant, their impact diminished in size and especially relative to the rapidly increasing effect of scale economies. By 1967, the contribution of technological changes dropped to only 4% of the "econometric" productivity gains. During the period 1968 to 1977, the share of the technology effect was fluctuating around 12%. As higher rates of introduction of new technology were registered and the scale effect declined somewhat due to the recent slowdown in demand for telephone services, the share of technological changes in "econometric" productivity gains increased to the 30% level in the last two years.

While the indexing approach to the measurement of Bell Canada's productivity produced an average annual gain of 3.5% for the entire period of observation, the average annual "econometric" gain is 3.68%. Technological progress in Bell Canada is directly responsible for a .88% average annual productivity improvement and scale economies generated a 2.80% average productivity gain per annum. Roughly one quarter of Bell Canada's productivity gains have been generated directly by technological changes and three quarters are due to the company's economies of scale. Technological changes are also the ultimate cause of a large part of the scale effect, because technological changes increased the degree of scale economies of Bell Canada and generated demand (hence scale increases) by lowering the cost of production, improving the quality and increasing the variety of telecommunications services.

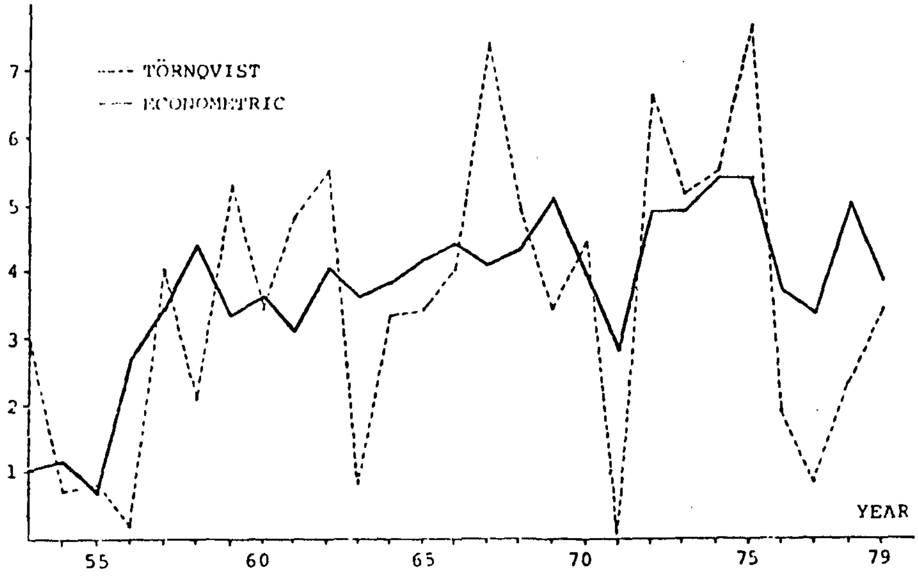


CHART 2: Annual Productivity Gains

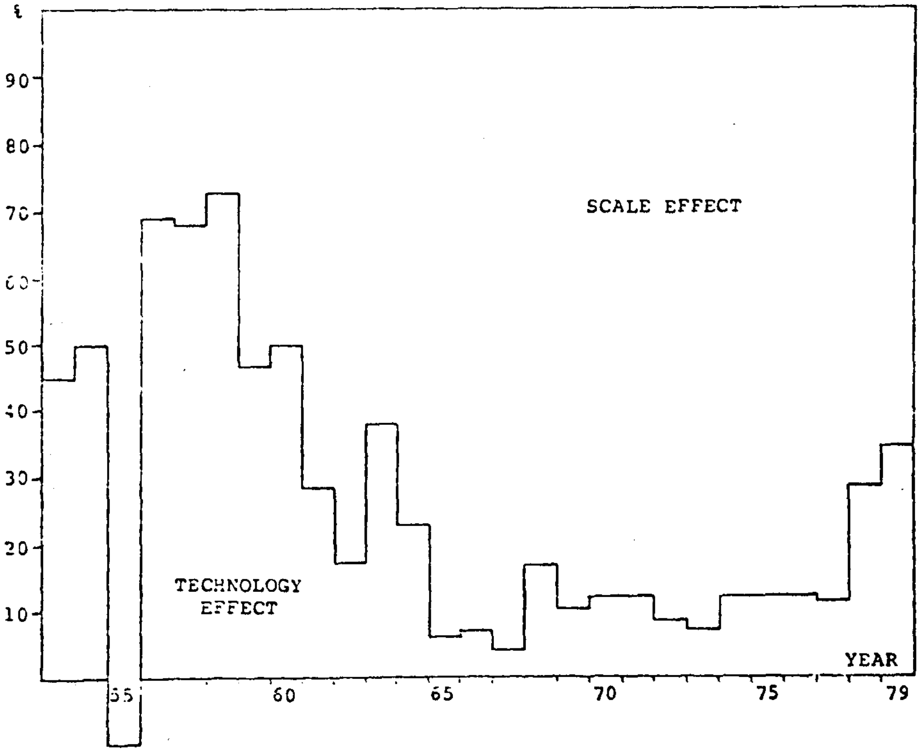


CHART 3: The Composition of Econometric Productivity Gains

FOOTNOTES

[1] *The two approaches are not separable. The verification of the validity of the notions of productivity utilized in the indexing approach requires econometric hypothesis testing and the econometric approach uses index numbers.*

[2] *The description generally follows Berndt (1980). The theory of duality is not explored here, but the econometric productivity gains are derived from a cost function as well as from a production function in sub-section 2.2.*

[3] $\dot{Q} = \frac{\partial Q}{\partial t} / Q; \quad \dot{X} = \frac{\partial X}{\partial t} / X$.

[4] *E.g., Solow (1957), Jorgenson and Griliches (1967).*

[5] *Denny, Fuss and Everson (1979), Denny, Fuss and Waverman (1979).*

[6] *Kiss, Karabadjian and Lefebvre (1981).*

[7] *Griliches (1963, 1964, 1967), Denny, Fuss and Everson (1979), Denny, Fuss and Waverman (1979), Nadiri and Schankerman (1979).*

[8] *Kendrick (1961).*

[9] *Fisher (1922), Törnqvist (1936), Theil (1967), Christensen and Jorgenson (1970), Diewert (1976, 1977), Jorgenson and Lau (1977).*

[10] *Diewert (1976).*

[11] *More testing is required before any definitive conclusion is drawn.*

[12] *The estimated single output translog cost model could not reject the hypothesis of homotheticity of technology. Samuelson and Swany (1974) and Usher (1974) showed that the Divisia index is path independent if the production function is homothetic; however, if constant returns to scale do not exist the Divisia formula does not yield the desirable index.*

[13] *Ohta (1974).*

FOOTNOTES (Cont'd)

[14] *Nadiri and Schankerman (1979) used a quasi-Divisia index to aggregate input; thus, equation (11) became*

$$(P\dot{R}) = -\dot{B}/\varepsilon_{CQ} + (k\varepsilon - 1)\dot{X} ,$$

where $k = \Sigma(P_i Q_i)/\Sigma(W_j X_j)$; *i.e., the revenue/cost or average price/average cost ratio.*

[15] *The alternative formula in equation (11) can be treated in a similar fashion.*

[16] *The technology effect is the residual of Törnqvist productivity gains in Denny, Fuss, Everson (1979) and Nadiri and Schankerman (1979). An alternative would be to estimate the technology effect from a cost function and attribute the other effects to the scale effect. However, this method would exhibit similar sensitivity to the error in the cost elasticity estimate. A further difficulty appears when X is used in the scale effect, because differences between the actual and cost minimizing input growth rates distort the residual technology effect.*

[17] *A separate residual term is shown in Denny, Fuss and Waverman (1979).*

[18] *Telephone exchange upgrouping is a reflection of qualitative improvement in local service output, resulting from increases in the size of the local calling area. Since the local service price index does not reflect rate increases are shown as output volume increases.*

[19] *E.G., third number and credit card calls.*

[20] *Smith and Corbo (1979), Denny, Fuss and Everson (1979).*

[21] *Gollop and Jorgenson (1980), p.67.*

[22] *Kendrick (1973), p.26.*

[23] *The term "quality" refers to the marginal revenue product of the factor in question. The marginal revenue product of labour is approximated by the hourly rate of total labour cost. The marginal rate of return or the user cost of capital.*

FOOTNOTES (Cont'd)

[24] *It is assumed that the 1978 estimates of cost elasticity with respect to output and technological changes prevailed in 1979 and the rate of increase in the technology proxy variable was 2.5% in 1979. The preliminary scale effect appears to be reliable, since the estimated output elasticities of cost were stable during the last years of the period of observation. However, the 1.3% estimate of the technology effect (see Table 4) may be overstated. E.g., if the technology elasticity of cost follows its tendency to decline and becomes -.57 (instead of -.53, as assumed), but the technology proxy grows only by 2%, the estimated technology effect becomes 1.13%.*

[25] *The Kendrick index with moving base year is approximately equal to the corresponding Tornqvist index in the long run. The differences are usually small in the sub-periods and there are a few larger deviations between the annual gain estimates (1956, 1959, 1966). The Kendrick index with moving base year has approximated the Tornqvist index reasonably well for Bell Canada.*

[26] *Tornqvist indices of productivity, computed in a very similar fashion for AT&T by Christensen, Cummings and Schoech (1980), make the following comparison of average annual gains possible:*

PERIOD	*AT&T*	*Bell Canada*
1957-66	*3.1%*	*3.7%*
1967-77	*3.2%*	*4.3%*

Denny, Fuss, and May (1980) reported average annual productivity gains in the .22 to 2.43% range for twenty two-digit manufacturing industries in Quebec and Ontario during the period 1961 to 1975. Bell Canada's productivity gains averaged 4.4% per annum in the same period.

FOOTNOTES (Cont'd)

[14] *Nadiri and Schankerman (1979) used a quasi-Divisia index to aggregate input; thus, equation (11) became*

$$(\dot{PR}) = -\dot{B}/\varepsilon_{CQ} + (k\varepsilon - 1)\dot{X} ,$$

where $k = \Sigma(P_iQ_i)/\Sigma(W_jX_j)$; *i.e., the revenue/cost or average price/ average cost ratio.*

[15] *The alternative formula in equation (11) can be treated in a similar fashion.*

[16] *The technology effect is the residual of Törnqvist productivity gains in Denny, Fuss, Everson (1979) and Nadiri and Schankerman (1979). An alternative would be to estimate the technology effect from a cost function and attribute the other effects to the scale effect. However, this method would exhibit similar sensitivity to the error in the cost elasticity estimate. A further difficulty appears when X is used in the scale effect, because differences between the actual and cost minimizing input growth rates distort the residual technology effect.*

[17] *A separate residual term is shown in Denny, Fuss and Waverman (1979).*

[18] *Telephone exchange upgrouping is a reflection of qualitative improvement in local service output, resulting from increases in the size of the local calling area. Since the local service price index does not reflect rate increases are shown as output volume increases.*

[19] *E.G., third number and credit card calls.*

[20] *Smith and Corbo (1979), Denny, Fuss and Everson (1979).*

[21] *Gollop and Jorgenson (1980), p.67.*

[22] *Kendrick (1973), p.26.*

[23] *The term "quality" refers to the marginal revenue product of the factor in question. The marginal revenue product of labour is approximated by the hourly rate of total labour cost. The marginal rate of return or the user cost of capital.*

FOOTNOTES (Cont'd)

[24] *It is assumed that the 1978 estimates of cost elasticity with respect to output and technological changes prevailed in 1979 and the rate of increase in the technology proxy variable was 2.5% in 1979. The preliminary scale effect appears to be reliable, since the estimated output elasticities of cost were stable during the last years of the period of observation. However, the 1.3% estimate of the technology effect (see Table 4) may be overstated. E.g., if the technology elasticity of cost follows its tendency to decline and becomes -.57 (instead of -.53, as assumed), but the technology proxy grows only by 2%, the estimated technology effect becomes 1.13%.*

[25] *The Kendrick index with moving base year is approximately equal to the corresponding Törnqvist index in the long run. The differences are usually small in the sub-periods and there are a few larger deviations between the annual gain estimates (1956, 1959, 1966). The Kendrick index with moving base year has approximated the Törnqvist index reasonably well for Bell Canada.*

[26] *Törnqvist indices of productivity, computed in a very similar fashion for AT&T by Christensen, Cummings and Schoech (1980), make the following comparison of average annual gains possible:*

PERIOD	*AT&T*	*Bell Canada*
1957-66	*3.1%*	*3.7%*
1967-77	*3.2%*	*4.3%*

Denny, Fuss, and May (1980) reported average annual productivity gains in the .22 to 2.43% range for twenty two-digit manufacturing industries in Quebec and Ontario during the period 1961 to 1975. Bell Canada's productivity gains averaged 4.4% per annum in the same period.

APPENDIX A: DATA DESCRIPTION

Appendix A: Data Description (Cont'd)

1. Output

Aggregate output volume changes are measured by Laspeyres and Törnqvist volume indices, while Paasche and Törnqvist price indices provide for the measurement of price changes in output aggregates.

1.1 Output Categories

The following ten output categories are distinguished in the Bell Canada productivity study:

1. Local services,
2. Intra-Bell message toll services,
3. Canada message toll services,
4. US and overseas message toll services,
5. WATS,
6. TWX,
7. Private line toll services,
8. Miscellaneous other toll services,
9. Directory advertising,
10. Miscellaneous services.

Their contents can be best described through the revenues they generate.

Local Service Revenues

Include contract basic charges (residence and business main and extension, PBX trunks and extensions and Centrex Co and Cu primary and secondary), contract auxiliary charges, non-recurring and message charges, local public telephone and private line revenues and other small items.

Intra-Bell Message Toll Revenues

Include all revenues derived from long distance calls originating and terminating in Bell Canada territory and some settled revenues from messages originated in independent telephone companies and terminated in Bell Canada territory as well as from Bell-originated and independent-company-terminated calls. These independent companies are located within or adjacent to Bell Canada territory.

Canada message toll revenues are, in theory, derived from long distance calls in both directions between Bell Canada and other member companies of the Trans-Canada Telecommunications System (TCTS). The settled revenues are internally estimated.

US and overseas message toll revenues are similarly derived from the two-way traffic between Bell Canada and foreign countries. Countries other than the United States are referred to as "overseas", regardless of their geographical location. The settled revenues are internally estimated.

WATS revenues originate from INWATS and OUTWATS services. OUTWATS permits customers to call and INWATS to receive calls from anywhere within or below specified "zones" for a flat monthly rate. Seven zones (reduced to six in October 1978) have been developed by forming concentric areas using the NPA divisions.* Rates vary by type, zone and home NPA. Rate types are:

- full line (unlimited or, after October 1978, a maximum of 160 hours calling time per month);
- measured line (maximum 10 hours calling time per month);
- half measured line (maximum 5 hours per month).

Calling time in excess of the allowed maximum (overtime) is billed at 80 to 85% of the hourly rate calculated at measured line rates. OUTWATS and INWATS have the same rate structure.

*NPA (Numbering Plan Area) is a 3-digit code used as a prefix identifying all telephone numbers within a defined geographical area. Also referred to as "area code".

For zones, where it is possible to call other TCTS company subscribers, the TCTS settlement plan deals with the WATS revenues of Bell Canada and of other member companies.

TWX revenues are derived from message charges and equipment rental charges. In addition to rental charges and intra-Bell message charges, this category also includes revenues from the TCTS settlement process.

Private line toll revenues originate from the sale of private line voice and data services; i.e., inter-exchange voice and teletypewriter private line, radio and TV program transmission, Telpak, Datapac, Dataroute and other data services. Again, total revenue includes revenues from the TCTS settlement for private line circuits that have one end-point in Bell Canada territory and the other end-point in TCTS member companies.

Miscellaneous other toll revenues are a residual category. The most important component services are Multicom and Voicecom.

Directory advertising revenues were derived from Yellow Pages advertising during the period 1952 to 1971. With the establishment of Tele-Direct as a Bell Canada subsidiary in 1971, this category was discontinued.

Miscellaneous revenues include Tele-Direct commission; rents of equipment, poles, buildings, satelite, etc.; general services and licences, e.g., service agreement revenues; Teleboutique/Phone Centre sales and various other revenues.

Although the productivity study uses a single output aggregate, for analytical purposes as well as for various econometric studies, some two-output and three-output subaggregates are also generated. The two-output subaggregates are defined as:

- local, directory and miscellaneous (Nos. 1, 9, 10)
- toll (Nos. 2, 3, 4, 5, 6, 7, 8);

and the three-output subaggregates are derived by breaking down the toll category in two ways into

A. - message toll (Nos. 2, 3, 4)
   - other toll (Nos. 5, 6, 7, 8) or

B. - "monopoly toll" (Nos. 2, 3, 4, 5)
   - "Competitive toll" (Nos. 6, 7, 8).

The relatively small directory advertising and miscellaneous service categories are aggregated with local services in order to minimize the consequences of (1) not having a price index for miscellaneous services, (2) a break in both series in 1971 due to the establishment of Tele-Direct; and also to reduce outputs to a manageable number in econometric studies. The "competitive toll" category is a rather crude approximation to the truly competitive toll services.

1.2 Output Prices

A Paasche or, alternatively, a Törnqvist price index should be assigned to each of the ten output categories. However, the available price indices do not always conform with the Paasche formula and Törnqvist indices are not available at all. The price indices generally have fixed volume weights in the early years (from 1952 to around 1970) and variable weights for the 1970's. Although the base periods and volume weights have been chosen in various ways, the fundamental features of the procedure of calculating the 1967-based index series shown in Table B.4 are common. When a change in telephone rates takes place, the appropriately chosen service volumes in the base year (q_{iB}) are priced out at old (p_{i0}) and new (p_{i1}) prices. The index formula (the "reprice" effect in Bell jargon)

$$P_{01} = \frac{\sum_i p_{i1}\, q_{iB}}{\sum_i p_{i0}\, q_{iB}} \quad (i=1, \ldots, n)$$

is applied, the year-to-year indices are chained,

$$P_{0t} = P_{01} \cdot P_{12} \cdot \; \ldots \; \cdot P_{t-1,t},$$

and the resulting index series is normalized around its 1967 value. This procedure yields general price levels, relative to 1967, for the n services included in the given output category. When price changes do not take place on January 1, the average annual level of the price index is calculated as the arithmetic mean of all price levels that exist in the given year, weighted by the portion of the year during which they exist. To simplify the process, the distribution of volumes through the year is not taken into account.

1.21 Local Price Index

For the periods 1952 to 1959 and 1969 to 1971, the fixed volume weights are those of December 1965. There were no price changes between 1959 and 1968. Paasche price indices have been used since 1972. Up to 1968, the indices cover residence and business main and extension telephones and PBX trunks and extensions, representing approximately 65% of total local service revenues. The coverage was extended to include non-recurring charges in 1969. 100% coverage was achieved in a detailed computer program in 1972. From an indexing point of view, it is an unusual feature that this computer program uses test year forecasts of service volumes as weights. Since the forecasts are usually fairly accurate, the resulting indices approximate Paasche price indices reasonably well. Rate changes due to the upgrouping of telephone exchanges are consistently ignored.

The local service price index is considered to be of good quality for most of the sample period. However, the index values in the 1952 to 1958 period are somewhat suspect, due to the partial coverage and the rather irrelevant 1965 volume weights.

1.22 Intra-Bell Message Toll Price Index

The index covers 100% of intra-Bell message toll services, but ignores - with negligible consequences - the revenue settlements with small independent telephone companies. Message volume weights are taken from a monthly sample of the traffic within and between Quebec and Ontario. The chosen representative months' weights are annualized with the aid of estimated seasonal factors. For the period 1952 to 1967, the base period was September 1967. October 1970 weights were used between 1968 and 1972, April 1972 weights between 1973 and 1977, and finally, June 1976 weights were utilized in the calculation of the rate change in 1978.

1.23 Canada Message Toll Price Index

Volume weights in the following base periods were used in the calculation of the index series:

- September	1967	(1953, 1959, 1960, 1962, 1964, 1966, 1968)
- April	1972	(1972)
- June	1974	(1975)
- June	1977	(1978)

The monthly sample data were annualized. Prices did not change in years which are not shown in brackets above.

The coverage of base year weights has undergone some changes since the initiation of the productivity study in 1968. For the September 1967 base period, all TCTS message toll traffic was included. Only calls affecting Bell Canada's settled revenues were included in the April 1972 base, i.e., the non-Bell adjacent member traffic was excluded. Finally, the forecast weights for 1975 and 1978 included the Bell Canada to adjacent member and Trans-Canada traffic only.

1.24 US and Overseas Message Toll Price Index

Volume weights in the following base periods were used to calculate the indices:

- September 1967 (1956, 1960, 1967, 1969)
- November 1974 (1975)
- June 1975 (1976)

Prices remained unchanged in years not shown in brackets above. All samples included the entire Bell Canada to US traffic.

Since a price index for overseas messages is not available, the US message toll service price index is used as a proxy for rate changes for overseas messages.

1.25 WATS Price Index

OUTWATS was introduced in February 1962 and INWATS in December 1969. A fixed-based (April 1972 weights) price index for Bell-originated OUTWATS contract charges in all zones (1 to 7) was constructed for the years 1964 to 1971. Beginning in 1972, the volume weights included Bell-originated INWATS and OUTWATS contract and overtime charges, but only in zones 1 to 4.

The quality of the WATS price index is considered adequate for the purpose of the productivity study. Productivity gains are insensitive to changes in the WATS price index. WATS revenues were included in miscellaneous other toll service measure in 1962 and 1963. This represents only a negligible source of error, since the initial service volumes were low. The less than full coverage of the volume weights does not create any major difficulty either, since there is sufficient reason to believe that there was little difference between changes in excluded and included prices. The only major weakness of the index lies in its fixed base year. The composition of WATS services has undergone significant changes, which should be reflected in the volume weights.

1.26 TWX Price Index

The index reflects changes in

- TWX equipment charges,
- Intra-Bell TWX message charges,
- Bell to US message charges.

Message charges to TCTS, adjacent member and overseas relations are excluded from the index. The fixed base year volumes refer to December 1972 for equipment charges and to April 1972 for message charges. Message volumes are based on a sample. Bell's own internal messages were excluded from the sample.

Intra-Bell message tariffs are based on originating and terminating NPA and on the duration of the message. Each NPA pair is converted into an approximate mileage and the minimun charge is ignored.

The index has several problems. Although TWX has not experienced rapid growth or major restructuring, the use of fixed volume weights might be a source of bias, especially with respect to messages. The coverage should be extended to include price changes in Bell to TCTS and overseas messages. The interim Bell to US rate increases, which have been in effect since October 1978, should also be reflected. Bell Canada equipment could not be excluded from the index; thus, a greater than desirable weight is given to price increases in equipment charges.

1.27 Private Line Toll Price Index

This is a very crude price index. The average circuit length of intra-Bell inter-exchange voice private lines in February 1973 (110.54 miles) is priced out at the different rates that existed during the entire sample period and the price indices are generated as ratios. The index ignores non-intra-Bell circuits, data and Telpak services and the changes in the mileage length composition of intra-Bell services. It is likely that the private line price index contains distortions.

1.28 Directory Advertising Price Index

The computation of this price index requires a special procedure as there are lags in the implementation of price changes, due to the fact that directories in different regions of Bell Canada territory are published at different times. The index series, which was discontinued in 1971, is not described here.

1.29 Miscellaneous Other Toll and Miscellaneous Service Price Indices

No price indices have been developed for the miscellaneous other toll, also referred to as "other other" toll or "data", and the miscellaneous service categories. Both are very heterogeneous and relatively small. In 1979, 3.9% of total revenue originated from miscellaneous services and only .5% from "other other" toll. The implicit price index (the ratio of current to constant 1967 dollar revenues) for TWX, WATS and private line toll is used as a proxy for "other other" toll and the implicit price index of local, message toll and other toll services is the proxy for miscellaneous services.

1.3 Output Volume

Output volumes are represented by deflated (constant dollar) revenues in the Bell Canada productivity study. The 1967-based index series of constant dollar revenues by output category are shown in Table B.1.

According to its contents, output is measured in three alternative ways as

- gross production,
- gross value added (gross production less materials),
- net value added (gross value added less depreciation).

Revenue-related non-income taxes can be used to adjust either the price or the volume of output. In Tables B.1 to B.3, the output volumes are adjusted. However, it should be noted that the revenue - related non-income taxes are so small in comparison to the total revenue of Bell Canada, that the difference between the two adjustments, as well as between the adjusted and unadjusted series, is negligible.

Laspeyres and Törnqvist volume indices are constructed for gross production, while a Laspeyres index formula is used for the two value added measures. Value added indices are not shown in the data tables, but Laspeyres volume indices of aggregate gross production and its sub-aggregates are given in Table B.2. The Laspeyres volume index is obtained as the ratio of constant dollar revenue to the base year's revenue. Correspondingly, the Paasche price index of gross production, shown in Table B.4, is the ratio between current and constant dollar revenues. Since volumes are represented by deflated revenues, the Q_{it}/Q_{it-1} individual growth rates of the Törnqvist volume index are replaced by the growth coefficients of constant dollar revenues; i.e., by the $(P_{iB}Q_{it})/(P_{iB}Q_{it-1})$ ratios. Törnqvist volume indices for gross production and its sub-aggregates are included in Table B.3. The implicit Törnqvist price indices of gross production and its sub-aggregates are obtained in Table B.5 by dividing the current dollar revenue index by the Törnqvist volume index.

The base year in the Laspeyres volume indices is either the traditional B=1967 or a moving B=t-1. In the latter case, and also in the case of Törnqvist indices, the annual indices are chained and the resulting series are normalized around 1967 to facilitate a comparison of index formulae. The Laspeyres volume indices of Table B.2 have 1967 as their fixed base year.

2. Labour Input

Changes in the quantity of labour input of Bell Canada are expressed by three volume index series in Table B.6. The first series contains unweighted indices of hours worked, while the other two series consist of Laspeyres and Törnqvist volume indices, in which hours worked within each labour category are weighted by their respective hourly total labour costs (Laspeyres) or their respective shares in total labour compensation (Törnqvist). The Laspeyres volume indices use 1967 as the fixed base year.

The three price index series in Table B.7 correspond to the volume indices of Table B.6. They show the changes in total labour cost per hour worked. The product of the volume index and its corresponding price index is equal to the index of total labour cost.

2.1 Labour Input Volume

Labour input is measured by the total number of hours actually worked by the Bell Canada labour force on the production of telecommunication services. Hours worked by occasional employees are now excluded from the calculations due to the relatively small number of employees (approximately .5% of total with .3% of total wages and salaries) and the high cost of collecting the information. Hours worked on the construction of telephone plant are also excluded.

Since qualitative differences among employees (based on occupation, education, experience, etc.) cause the marginal products of their labour input to vary, the need for a homogeneous labour input category necessitates an adjustment to hours worked. Under the conventional assumption that, as in the case of competitive equilibrium, the marginal product of labour is equal to its rate of compensation, an index of manhours worked is obtained by using weights based on the wage rates in each available labour category.

Quality-adjusted manhours are not directly available from Bell Canada records. They are estimated through the following five steps of calculations:

1. Annual average number of employees by category.
2. Average hours worked per employee per year by category.
3. Total unadjusted hours worked.
4. Total unadjusted expensed hours worked.
5. Quality adjustment of expensed hours worked, by category and total.

2.11 Annual Average Number of Employees

Bell Canada employees can be classified as:

- regular full-time,
- regular part-time,
- temporary full-time,
- temporary part-time,
- occasional.

A regular employee is an employee whose employment is reasonably expected to continue for longer than one year, although such employment may be terminated earlier by action on the part of the company or the employee. Temporary employees are engaged on the understanding that the period of employment is expected to continue for more than three weeks but not more than one year. Occasional employees are engaged for periods expected to last less than three consecutive weeks. A full-time employee is normally required to work the basic hours of work and part-time employees work less than the basic hours.

The following six occupational groups are distinguished with respect to regular and temporary full-time employees:

1. Telephone operators
2. Plant craftsmen
3. Clerical
4. Other non-management
5. Foremen and supervisors
6. Other management

Since the marginal product of labour is assumed to increase with experience, the quality adjustment of hours worked requires that each occupational group be disaggregated into sub-groups according to the length of service. Different occupational groups have different length-of-service distribution and the disaggregation has been made in such a way as to have approximately the same number of employees in each sub-group. For the six occupational groups, there is a total of 26 sub-groups. Occupational groups and sub-groups are not distinguished for part-time employees, due to their relatively small number.

The year-end number of employees is available from Bell Canada records. The annual average number of employees is taken as the simple arithmetic mean of the number of employees at two consecutive year-ends. The averages are summed up for each occupational group and then for full-time and part-time employees in order to arrive at the annual average number of total employees. The number so obtained is based on the assumption that the path of the number of employees is linear between year-ends and it leads to a downward bias, due to seasonal variation in the number of employees. The typical seasonal pattern shows local maxima around mid-year and local minima at year-end. The difference between the 12 month average and the year-end average is prorated among the occupational groups of full-time employees in each year. For part-time employees, it is more realistic to assume a linear path between year-ends than to assume a path similar to that of full-time employees.

2.12 Average Hours Worked per Employee per Year

2.121 Full-time Employees

Hours worked are obtained as scheduled (basic) hours, minus the hours of vacation days, scheduled days off (SDO's), holidays and sickness leave, plus overtime hours worked.

Scheduled hours differ among, and also within, occupational groups as well as according to the length of service, depending on Bell's policy for management employees and the collective agreements for non-management. When more than one collective agreement, containing different numbers of scheduled hours, applies for a year, or employees in an occupational group come under different collective agreements, then the scheduled hours are calculated as a weighted average of the different hours and the average is rounded to the nearest quarter of an hour. The weights are the portions of the year during which the different agreements are in effect or the number of employees under different agreements within each occupational group or sub-group.

Vacation days are determined by the Company's policy for management employees and by collective agreements for non-management. Vacation hours are calculated at the scheduled hours per day for each sub-group. It is assumed that the total number of granted vacation days is taken by all employees. Employees with less than one year of service are generally granted one vacation day per month of service up to a maximum of ten days. The average number of vacation days in the 0 to 6 months service sub-group is established at 3 and it is 8 days for employees with 6 to 12 months of service.

Holiday hours are arrived at by multiplying the number of holidays by the number of scheduled hours per day. Together with one day off with pay every year, granted by Bell Canada to all its employees, the total number of holidays is now ten. The number of scheduled days off (SDO's) is determined by company policies and collective agreements. There are other losses in worked hours (e.g., coffee breaks, union meeting, bereavement, jury or witness duty, election, etc.) which are not subtracted from scheduled hours.

Data are available for overtime payments but not for overtime hours. Overtime payment is equal to the number of overtime hours times the basic wage or salary rate, increased by a multiplicative factor, which is usually 1.5. Thus, overtime hours are calculated from data on overtime payments and the basic wage and salary rates.

2.122 Part-Time Employees

The majority of part-time employees are concentrated in two occupational groups: telephone operators and clerical. Average hours worked per employee per year are calculated for these two groups and the other groups are ignored. More accurately, the number of part-time clerical employees is calculated as the difference between total part-time employees and part-time telephone operators, i.e., it includes part-time employees in all other occupational groups.

It is assumed that part-time employees work 48% of the hours worked by their full-time counterparts. Average hours worked per part-time employee per year so derived are finally weighted together with the ratios of part-time employees in each group to total part-time employees.

2.13 Total Unadjusted Hours Worked

The annual average number of employees is multiplied by the average number of hours worked per employee per year in each of the 26 occupational sub-groups for full-time employees and also for total part-time employees in order to get the number of total unadjusted hours worked. Both expensed (worked on the production of telecommunication services) and capitalized (worked on the construction of telephone plant) hours are included in this measure.

2.14 Total Unadjusted Expensed Hours Worked

Data on expensed or capitalized hours worked are not directly available. The calculations are done in several steps.

1. The ratio of construction employees to total employees by departmental group (c_j) is calculated. The ratio is not available for occupational groups.

2. A transformation matrix is constructed with general element e_{ij} showing the average (of two year-ends) number of employees in occupational group i and accounting group j.

3. The average number of non-construction employees by occupational group is arrived at as

$$e_i^c = \sum_j (1 - c_j)\, e_{ij}.$$

4. The ratio of non-construction to total employees by occupational group is given as

$$x_i = e_i^c/e_i,$$

where $e_i = \sum_j e_{ij}$.

5. Expensed manhours by occupational sub-group are

$$h_{ik} = x_i h'_{ik},$$

where h'_{ik} is the number of total unadjusted hours worked in occupational group i and sub-group k.

6. Total unadjusted expensed hours worked are simply summed up for occupational groups, $h_i = \sum_k h_{ik}$ and for Bell Canada, $LU = \sum_i h_i$.

The departmental groups considered are:

1. General Offices,
2. Engineering,
3. Commercial and Marketing,
4. Plant,
5. Traffic.

The ratio of construction employees in General Offices (j=1) is approximated by the ratio of expenses charged to construction to total General Offices expenses. Expenses include wages, salaries, fringe benefits and other expenses. The expenses are collected from the appropriate Bell Canada expense accounts and the ratio is manually calculated. In Engineering (j=2), the ratio of construction expenses to total engineering expenses is manually calculated from data obtained from internal Bell Canada accounting reports. There is no construction activity in Commercial and Marketing and in Traffic; thus $c_3=c_5=0$. For plant employees (j=4), the ratio of construction employees to total employees is taken. The number of total employees is readily available and the number of construction employees is calculated as the ratio between total wage payments charged to construction and the average wage per year for construction employees. The latter is not available and had to be approximated by dividing total wage payments (excluding occasionals) by the average number of plant employees. The approximation involves the assumption of equal wage rates for construction and non-construction employees.

The transformation matrix covers all full-time employees. Its elements are determined on the basis of job duty codes.

For part-time employees, the above described procedure is not followed but the ratio of non-construction to total employees is calculated directly as the weighted average of the same ratios for telephone operators (x_1) and clerical (x_3), where the weights are the percentages of total part-time employees in each respective occupational group.

The x_i ratios are applied to total manhours under the assumption that the number of hours worked by construction employees and non-construction employees is the same in each occupational group; thus, the percentage of manhours in construction coincides with the percentage of employees in construction. It is also assumed that the x_i ratio is the same for each sub-group within occupational groups, i.e., it does not vary with the length of service.

2.15 Quality Adjustment

Quality-adjusted total expenses manhours (LA) are obtained by multiplying any arbitrarily chosen base year's unadjusted total expensed manhours by a labour volume index which refers to the same base year. Two types of volume indices are used:

1. A Laspeyres volume index with 1967 as the base year.
2. A Törnqvist volume index.

The weights of the labour volume index formulae are wage rates (Laspeyres) or compensation shares (Törnqvist). The disaggregated wage rates are described in Sections 2.22 and 2.23 of this appendix.

2. Labour Input Price

2.21 Average Labour Price

The average labour price is the hourly rate of total actual labour expense: Employee Expense (wages, salaries, fringe benefits) and five labour-related federal and provincial taxes (Canada Pension Plan, Quebec Pension Plan, Unemployment Insurance, Quebec Health Insurance Plan, Workmen's Compensation). Employee Expense includes only the expensed portion of wages, salaries and fringe benefits, but labour-related other taxes (LROT) include capitalized as well as expensed tax items. Capitalized labour-related other tax (CAPTAX) is subtracted from LROT before the latter is added to other labour expenses.

The average labour price is

$$\overline{W} = \frac{EE + (LROT - CAPTAX)}{L}$$

Where EE denotes employee expense and L signifies either adjusted (LA) or unadjusted (LU) hours worked.

2.22 Disaggregated Labour Prices (1966-1979)

Disaggregation by occupational groups and sub-groups matches that of hours worked. Disaggregated labour prices contain the same items as the aggregate labour price.

Wage and salary data are collected, together with the number of employees. Year-end levels of weekly wages and salaries per employee are annualized by multiplying them by 52.2. The annualized wage and salary rates are divided by the average number of hours worked per employee per year in order to get the wage portion of the disaggregated labour price for each labour sub-group.

The disaggregated labour price in each sub-group is the sum of its wage component and fringe benefit component. The latter is taken into account as follows.

Only the total cost of fringe benefits and labour-related other taxes is available from Bell Canada records. Some of the reported costs are included in the wage rate. After deducting the cost of paid leave benefits, grievances and negotiations, disability pension and scheduled days off and adding accident disability expenses, the total cost of fringe benefits not included in the basic wage rate is obtained.

The following benefits are paid by the company for temporary part-time employees: government pension plan, medical facilities, workmen's compensation, unemployment insurance. The total cost of these benefits is obtained from Bell Canada records and is distributed between temporary part-time and other employees according to their respective shares in the Company's total wage payment. After removing the cost of benefits for temporary part-time employees, the remainder is prorated among occupational groups, according to their shares in total wage payment. Within occupational groups, the prorating is done according to the average number of employees in each sub-group; i.e., it is assumed that the cost of benefits per employee is proportional to wage rate among occupational groups but it is insensitive to wage rate differentials within the same occupational group. In each sub-group, the total cost of fringe benefits is divided by the average number of employees and the resulting fringe cost per employee is further divided by the average number of hours worked per employee per year in order to obtain the fringe benefit portion of the disaggregated labour price.
The number of part-time employees is broken down into regular and temporary employees, but their fringe benefit per employee values

are averaged, because only an average number is available for hours worked per employee per year.

Table B.9 contains the disaggregated labour input prices for the 1967 to 1979 period.

2.23 Disaggregated Labour Prices (1952-67)

The Bell Canada records from which information on disaggregated labour input prices originates have not been preserved for years before 1967. The only available source is a special study providing total wage payments for the following three occupational groups:

1. telephone operators,
2. plant craftsmen,
3. clerical employees.

For each occupational group and each year, total wage payment is divided by the average number of employees and further divided by the average number of hours worked by an employee in order to obtain hourly wage rates. The time series of wage rates are normalized around 1967 and the resulting index series are multiplied by the 1967 labour price, described below.

A fourth labour class is generated as the residual of total wage payment and employees over the sum of the three occupational groups, available from the special study. Indices (1967=1.0) of hourly wage rates are derived as for the other three categories and are multiplied by the 1967 labour price of the residual category. Labour prices in the residual class may be slightly distorted. It is not known whether the special study treated wage payments exactly in the same manner as they were treated in the regular accounting reports. A careful evaluation and comparison found the disaggregated labour prices realistic for the 1952 to 1967 period.

The 1967 labour prices are calculated in the table below. Wages for operators, craftsmen and clerical employees are taken from the special study. Total wages originate from regular internal reports and wages of other employees are taken as the residual. Expensed ratios for operators, craftsmen and clerks are from the labour volume calculations (see Section 2.14 above). Expensed wage is the product of total wage and the expensed ratio for the three occupational groups. Total expensed wage is obtained from internal reports and the fourth labour class is taken again as the residual. Total labour cost is the same as the numerator of the aggregate labour price formula in Section 2.21 above. Expensed wage in each of the four labour classes is multiplied by the ratio of total labour cost to total expensed wage to get labour cost by class. Disaggregated labour prices in the last column of the table are obtained by dividing labour cost by the number of expensed manhours worked.

Disaggregated Labour Prices, 1967

Labour Category	Wage Payment (000)	Expensed Ratio	Expensed Wage (000)	Labour Cost (000)	Expensed Manhours (000)	Labour Price ($/Hour)
Operators	25,916	1.0000	25,916	28,026	12,362	2.27
Craftsmen	63,832	.6378	40,712	44,027	12,902	3.41
Clerical	36,118	.7792	28,143	30,434	11,828	2.57
Others	110,915	.7786	86,358	93,389	19,488	4.79
TOTAL	236,781	.7650	181,129	195,877	56,580	3.46

Table B.8 contains the disaggregated labour prices for the period 1952 to 1967. The weighted labour input volume indices in Table B.6 were calculated with these prices for the period 1952 to 1967 and with the more disaggreated and more reliable labour prices shown in Table B.9 for the period 1967 to 1979.

3. Capital Input

The stock of capital is used as an approximation to capital input. The following three alternative definitions have been considered in the Bell Canada productivity study:

1. plant in service,
2. plant in service and under construction,
3. plant in service and under construction, plus working capital.

The first two definitions refer to the stock of physical capital. As Table B.10 reveals, there are only negligible differences between the volume indices of the two measures, because the volume of plant under construction (PUC) is very small in comparison with the volume of plant in service. The third definition is not explored in this paper.

3.1 Capital Input Volume

Plant in service includes land and depreciable plant. It consists of the following six major categories:

1. Land,
2. Buildings,
3. Central office equipment,
4. Outside plant,
5. Station equipment,
6. General equipment.

The second definition of capital stock adds a seventh major category: plant under construction (PUC).

Capital input volumes in each category are represented by constant dollar stocks of physical capital. These are obtained from book values by restating their age distribution by appropriately constructed price indices. An unweighted index of constant dollar net stocks is calculated, together with Törnqvist volume indices, in Table B.10. The Törnqvist index uses category shares of total capital compensation as determined by the residual rates of return as well as by the user costs of capital (see section 3.2 in this appendix). Gross capital series are not shown in Appendix B.

3.11 The Annual Average Stock of Physical Capital (Plant in Service)

The calculations of the stock of physical capital require the following information for each year.

1. BG_{ij}, the year-end book value gross plant in service in plant category i and vintage group j.
2. RES_{ij}, the estimated depreciation reserve in plant category i and vintage group j.
3. TPI_{ij}, the Telephone Plant Price Index (1971=100) in class i for year j.

The age distribution of gross plant in service is obtained by approximately 75 categories from the Bell Canada depreciation study. The vintage groups generally go back to 1920. Estimated reserves by category (calculated according to the ELG (Equal Life Group) method, including Bell Canada adjustment procedures) are added up to the account level and are balanced against the accumulated depreciation on each account. The account level actual/estimated reserve ratio is applied to estimated reserves in each component category. Net plant is calculated as

$$BN_{ij} = BG_{ij} - RES_{ij}.$$

The Telephone Plant Price Indexes yield 'translators' which show the rate of change in purchase price level in category i from year j (the year of the purchase) to any arbitrarily chosen base year (c); i.e.,

$$P_{ij} = \frac{TPI_{ic}}{TPI_{ij}} \quad .$$

Gross and net telephone plant in service in base year dollars by category is given as

$$KG_i = \sum_j P_{ij} \; BG_{ij}$$

and

$$KN_i = \sum_j P_{ij} \; BN_{ij} \quad .$$

Plant in service in each category is restated into 1967 dollars and also into current dollars in each year of the observation period. When a different base year is desirable, the current dollar plant value is restated to the required year's dollars by calculated price indices at the major plant class level. E.g., plant in t-1 dollars is calculated for the moving base year Kendrick index of productivity as follows. Two implicit price indices are obtained as

$$K_{t-1}^{(t-1)} \;/\; K_{1967}^{(t-1)} = P_{67,t-1}$$

$$K_{t}^{(t)} \;/\; K_{1967}^{(t)} = P_{67,t}$$

where the superscripts refer to the year at the end of which plant in service is measured, the subscripts refer to the year to the purchase price level of which plant is restated and references to gross or net value and to major plant class have been omitted for simplicity.

The required t-1 dollar value of plant in service at the end of year t is given as

$$K_{t-1}^{(t)} = K_t^{(t)} / (\frac{P_{67,t}}{P_{67,t-1}}) .$$

Gross and net plant by major class are obtained by adding up the restated (1967 and current dollar) values of plant in service in the component categories. The resulting values are related to the end of each year. Annual averages by major class are calculated as simple arithmetic means of two consecutive year-end figures for both constant (1967) and current dollar values. Implicit price indices are taken as ratios of current to constant dollar average plant in each major class and in each year. Due to the averaging, the implicit indices are not equal to one in the base year (1967). They are normalized around their 1967 value and the constant (1967) dollar average plant in service is obtained by dividing the current dollar average plant by the normalized implicit price index in each major class and in each year.

Price indices and age distribution are not available for land. It is assumed that land has the same age distribution and is subject to the same price changes as buildings.

The above described calculations are done by computer for 1975 and subsequent years. Full information is not available for years prior to 1975; therefore, the procedure had to be changed. For the period 1970 to 1974, the age distribution of plant and reserve were determined at the category level, using the procedure outlined above. However, no summaries were made at the major plant class level and the input data files are no longer available, making reruns extremely costly. Category age distributions of plant at original cost were generated for major classes manually, using a sample of equipment to reduce the work load. Representative items of major classes, covering over 75% of the classes (with the exception of General Equipment where 61%

is covered), were selected, their estimated BG_{ij} and RES_{ij} values were raised to the level of total coverage and the restatements were carried out by special composite telephone plant indexes constructed for major plant classes. This method has the advantage of using information developed for the full plant age distribution study.

Further changes were necessary for the period 1952 to 1969, because the plant age distribution study had not been done for these years. Annual gross additions were collected from accounting records for each year beginning in 1920 (earlier gross additions were included in the 1920 value) and were summarized into major plant classes. A_{ij} values, signifying gross additions to plant in class i and in year j resulted. Survivor curves were selected which were considered to be representative of life and retirement dispersion for each class of plant in different periods of time. Survival ratios S_{ijt} were derived for each class i for plant purchased in year j, showing the ratio of plant surviving until the end of year t. At the end of any year t, the stock of gross capital was computed as

$$KG_{it} = \sum_{j=1920}^{t} S_{ijt} A_{ij} ,$$

with S_{ijt} A_{ij} being the age group for purchase year j. KG_{it} was balanced to the company books by a method called "computed mortality" which makes minor adjustments to the assumed life of equipment until KG_{it} is equal to the book value. The selected survivor curves also served as the basis for calculating reserve ratios which in turn were used to estimate reserve by plant class. Estimated reserves were balanced against actual accumulated depreciation. As in the full plant age distribution study, net plant is equal to gross plant minus estimated reserve. Finally, special composite telephone price indices for each major plant class were developed to facilitate the re-pricing of book value gross and net plant in a fashion identical to that of the full study.

There are some limitations in the above described process. The age distributions are estimated and may not correspond exactly to the actual age distributions. The translators and the survivor curves are composites for several depreciation categories and are not as accurate as those used at the category level.

Results for the periods 1952-1970 and 1970-1975 were reviewed for reasonableness and consistency with the full plant age distribution study. The calculations were carried forward to 1976 so that comparisons with the full study could be made. The simplified methods have been found to approximate the results of the full study reasonably well.

3.12 Plant under Construction

Plant under construction (PUC) at the end of each year is the sum of total telecommunications property under construction and telephone plant acquisition adjustment. Average PUC is calculated as the simple arithmetic mean of PUC at the end of two consecutive years. The constant dollar value is calculated by deflating current dollar average PUC by a composite telephone plant price index which shows the rate of change in purchase price level between the current year and the base year.

3.13 Working Capital

Working capital is defined as current assets minus current liabilities on Bell Canada's balance sheet, plus inventory at the end of the year. In order to better facilitate restatement, the difference between current assets and current liabilities is broken down to cash plus short-term deposits and net receivables (the difference between total accounts receivable and total accounts payable). The data are available from internal accounting reports. No price indices have been built for the components of working capital. Repricing to 1967 or any other year's dollars is done with the composite TPI for inventories, the GNE Implicit Price Index of Statistics Canada for cash and short-term deposits and the implicit price index of

Bell Canada's total revenue for net receivables.

3.2 Capital Price

Capital price is the cost to Bell Canada of a unit of its physical or extended capital. The total capital cost is determined by multiplying the stock of capital either by the residual rate of return or by the user cost in each major category and the indices of capital price are calculated in Table B.11 as implicit price indices by dividing the total capital cost index by the volume index of capital. The table contains net capital price indices only.

3.21 The Residual Rate of Return

Revenue is equal to total cost (including normal profit),i.e., payments to the factors of production for their productive services fully exhaust the firm's output, in the case of competitive equilibrium or when output prices are set according to their average costs (average cost pricing). Assuming output exhaustion in each year, the cost of capital is set equal to a part of the output value, which is not paid out to other productive factors (labour and material). The residual return is then divided by the stock of physical capital,

$$k_t = \frac{P_t Q_t - w_t L_t - m_t M_t}{K_t},$$

where w, L, m and M denote the prices and volumes of labour and material respectively. Each productivity measure has its own unique k_t residual rate of return to capital, which depends on the applied output and capital volume measures.

The residual rate of return can be calculated as the sum of capital-related costs per unit of capital. In the standard three-input model, capital-related costs are

- Depreciation (DEP),
- Capital-related non-income taxes (CROT),
- Income tax (TAX),
- Interest charges, excluding interest charged to construction (INT),
- Net income (NI).

Depreciation is obtained from accounting records. CROT is the sum of Quebec Capital Tax, Miscellaneous Provincial Taxes, Miscellaneous Federal Taxes and Property and Business Tax. Income tax and interest charges are obtained from the Bell Canada income statement. Net income is re-defined for the purpose of the residual rate of return calculations. Other income, unrealized foreign exchange gains or losses, contract operations and extraordinary items are included in the re-defined net income. The aggregate residual rate of return on capital is

$$k_t = \frac{DEP_t + CROT_t + TAX_t + INT_t + NI_t}{K_t} .$$

Disaggregated residual rates of return for the six major plant classes are obtained from the formula

$$k_{it} = \frac{\left(NI_t + TAX_t + INT_t + CROT_t\right)\frac{KC_{it}}{\Sigma KC_{it}} + DEP_{it}}{K_{it}}$$

where KC denotes current dollar capital stock, K denotes constant dollar capital stock, and subscript i refers to plant class.

An arbitrary solution to the measurement of category level residual rates of return is necessitated, because the residual returns are not observable at the category level. The solution adopted in the productivity study is simple pro-rating of the total residual return to capital among plant classes. The percentages of current dollar capital stock were chosen arbitrarily over those of book value or constant dollar capital stock, mainly because they allow the prorated return per unit of physical capital to vary in response to differences among plant classes in the rate of price changes between the base year and the current year. The prorated part of the rate of return is higher/lower than the average if the increase in purchase price level is higher/lower than the average.

3.22 The User Cost of Capital

The assumption of output exhaustion is relaxed through the use of the user cost of capital. Economic profit/loss exists if the user cost is less/greater than the residual rate of return. Although regulation supposedly prevents Bell Canada from incurring significant amounts of profit or loss in the long run, short-run non-zero profits may well exist as a result of errors in regulation (due to imperfect information, forecasting errors, etc.).

Various expressions of the user cost have appeared in the literature; e.g., Jorgenson (1963, 1967), Hall and Jorgenson (1967), Christensen and Jorgenson (1969), Boadway and Bruce (1979), Fuss and Waverman (1980), and Boadway (1980). All measures have been derived from the neoclassical theory of capital accumulation. Differences in the user cost measures arise from variation in the assumptions made in the formulation of the investment problem of the firm.

Cost minimizing behaviour is assumed for Bell Canada. In order to minimize the production cost of a given level of output, the company accumulates physical capital until the price of capital equals its marginal product times the marginal cost of producing the given level of output. Following Boadway (1980), the investment problem can be restated in a dynamic context as one of accumulating physical capital until the unit cost of physical capital equals the present value of the marginal cost of production times the marginal product of capital services.

At any point in time (t), the user cost of capital can be calculated from a perpetual inventory of capital stock model, where (in the presence of income taxes) the purchase price of capital goods (q_t) is

$$q_t = \int_t^{\infty} e^{-r(s-t)} \left[(1-u)c_s f_s e^{-\delta(s-t)} + uq_t D_{s-t} \right] ds \ ,$$

where r denotes the discount rate; c_s is the marginal cost (excluding depreciation allowances) of production at time s; f_s refers to the marginal product at time s of the stock of physical capital accumulated at time t; δ is the rate of economic depreciation; $e^{-\delta(s-t)}$ shows the rate at which the marginal product of capital accumulated at time t deteriorates by time s due to economic depreciation; u is the corporate income tax rate; finally, D_{s-t} is the depreciation for tax purposes at time s, relative to the original cost of physical capital of age s-t.

Denoting the present value of future depreciation deductions for tax purposes allowed on $1 current investment by z, the equation can be re-written as:

$$q_t = \int_t^{\infty} e^{-r(s-t)} \ (1-u) \ w_s e^{-\delta(s-t)} \ ds + uq_t z \ ,$$

where $w_s = c_s f_s$ is the user cost or the cost of capital services at time s.

Differentiating the re-written equation with respect to t and solving for w_t yields

$$w_t = [q_t(r+\delta) - \dot{q}_t] \frac{1-uz}{1-u} \ .$$

The $\dot{q}_t$ term represents capital gain from the resale of telephone plant. Its value is assumed to be zero*. w_t then becomes

$$w_t = q_t \ (r+\delta) \ \frac{1-uz}{1-u} \ .$$

*The annual average productivity gain for the 1953 to 1979 period increases from 3.52% to 3.64% if this assumption is relaxed.

This expression is analogous to the rental price of capital formula of Hall and Jorgenson (1967) (if tax credit is not assumed).

The user cost of capital is measured for each of the six major classes of physical capital. For classes of depreciable plant, the applied formula is

$$w_{it} = q_{it}(r_t+\delta_{it}) \left(\frac{1-uz_t}{1-u_t}\right) + \frac{CROT_t}{KN_t} \qquad (i=2, \ldots 6)$$

and for land it is written as

$$w_{it} = \frac{q_{it}}{1-u_t} r_t + \frac{CROT_t}{KN_t} , \qquad (i=1)$$

where q_{it} is the price of physical capital in category i, measured by an implicit* TPI_{it}; δ_{it} and u_t are as defined above; KN_t is the total net stock of physical capital in constant 1967 dollars and $CROT_t$ refers to the sum of capital-related non-income taxes as described in the preceding section (3.21). The measurement of the discount rate r_t and the present value of depreciation for tax purposes z_t is described below.

The discount rate r_t is the weighted average of the cost of new long term debt and equity capital; i.e.;

$$r_t = d_t(1-u_t)\ DRATIO_t + e_t(1-DRATIO_t) ,$$

where d_t is the cost of new long term debt; $DRATIO_t$ is Bell Canada's debt ratio (debt/debt plus equity); e_t is the expected rate of return on common equity. For simplicity, the relatively small amount of preferred equity is assumed to have the same rate of return as e.

*I.e., weighted by volumes of plant in service, as opposed to gross additions.

The expected rate of return on common equity is approximated by the expression

$$e_t = \frac{D_t}{P_t} + G_t ,$$

where D_t is common dividends declared per common share; P_t is the annual average market price of the common stock; and G_t is approximated by the 10-year log-linear least squares growth rate of dividends per share.

z_t depends on the method of depreciation deductions. Bell Canada followed the straight line depreciation method for tax purposes for the years 1952, 1953 and 1958 to 1966, and the declining balance method in other years. For the straight line depreciation, z_t is calculated as

$$z_t = \frac{1}{nr_t} [1 - \frac{1}{(1+r_t)^n}] ,$$

where n is the average life of the depreciable asset, approximated by taking the arithmetic mean of the reciprocals of composite book depreciation rates for each year between 1958 and 1966. Under the declining balance method, z_t is calculated as

$$z_t = \frac{CCA_t}{r_t + CCA_t}$$

where CCA_t is the composite federal cost of capital allowance rate, obtained from internal sources.

4. Material Input Volume and Price

An aggregate of miscellaneous inputs is referred to as material input in the productivity study. Material input is broken down into the following nine categories:

1. Maintenance material (all non-labour expenses related to maintenance of station equipment, COE and outside plant done by Bell Canada);
2. Contract maintenance (all expenses related to maintenance of station, COE, OP and buildings by contractors such as Northern Telecom);
3. Vehicles and tools (including gasoline expenses and vehicle rentals);
4. Rentals (real estate, circuits, poles, computers, etc.)
5. House service (electricity, fuel oil, other supplies, contract services, etc.)
6. Postage, printing, stationery;
7. Travel and transfer;
8. Research and development (mostly external, e.g., BNR);
9. Miscellaneous (e.g., advertising).

The only material-related non-income tax, the Ontario Official Telephone Service Tax, which has been levied since 1967 in lieu of retail sales tax on the telephone services Bell provides for its own use, is considered in the miscellaneous category.

The total expense associated with material inputs is greater than the Other Expense item of Bell's accounting reports in and after 1972, because an upward adjustment is made to eliminate the effect of capitalizing leased plant. Plant leases were originally included in Other Expense. In 1977, a change was made in the accounting procedures. Plant leases were excluded from Other Expense and leased plant values were included in telephone plant. A retroactive adjustment was possible only to 1972. In order to maintain the consistency of the capital depreciation and Other Expense time series, it was decided that the accounting change should be ignored and an adjustment to maintain the original method should be made in each subsequent year.

The breakdown of Other Expense is available only for 1971, 1972 and 1973 (but will be made available on an annual basis in the future). Since the distribution is not expected to change significantly, the average percentages for 1971 to 1973 were calculated for all component categories and the total adjusted Other Expense was prorated in each year from 1969 to 1979 on the basis of those average percentages in order to obtain the breakdown of Other Expense. No breakdown is used for the period 1952 to 1968.

Material price is represented by price indices. No price index has been developed for the period 1952 to 1968; thus, the GNE Implicit Price Index of Statistics Canada was used as a proxy for the composite price index of all components. Internally developed price indices are available for each of the nine categories from 1969 onward. The individual prices are observed in a number of internal sources (e.g., purchase accounting, contracts) and various price indices from Statistics Canada are used whenever internal prices are not observable. Approximately 80% of the prices are specific to Bell Canada.

The nine individual price indices are used to deflate the corresponding current dollar costs. The resulting constant dollar costs represent individual volumes. Volume and price aggregation for the period 1969 to 1979 is achieved by calculating

- Laspeyres volume indices and implicit Paasche price indices;
- Törnqvist volume and (implicit) price indices.

These indices are linked to the GNE deflator and to the index of constant (1967) dollar material costs to obtain the series for the entire sample period in Table B.12.

5. Depreciation

The Bell Canada productivity study requires major plant class level depreciation rates.

In Bell Canada, individual depreciation rates for 35 categories of telephone plant are established by the company's depreciation experts. These rates are applied to the corresponding categories of average (12-month simple arithmetic mean) book value depreciable gross plant in order to get annual book value depreciation expenses. The 35 depreciation expenses are summed up according to the five major classes of depreciable gross plant. Book value average depreciable plant is summed up in an identical fashion and the major plant class level depreciation rates, referred to in Section 3.2 of this appendix, are obtained as

$$\delta_i = \frac{\sum_d \delta_{id} \, BG_{id}}{\sum_d BG_{id}} \quad (d=1, \ldots, 35),$$

where d refers to the number of depreciation categories in plant class i, δ denotes the depreciation rate and BG is average book value depreciable plant. Table 13 of Appendix B contains the depreciation rates.

APPENDIX B: DATA TABLES

TABLE B.1: OUTPUT VOLUME INDICES BY CATEGORY

YEAR	LOCAL	INTRA-BELL	TRANS-CANADA	US, OVERSEAS	WATS	TWX	PRIVATE LINE	MISC. OTHER TOLL	MISC., DIRECTORY
1952	30.83	29.56	9.54	15.68			5.03	4.55	38.02
1953	33.41	31.61	10.71	17.83			6.93	5.22	43.03
1954	36.09	33.84	11.93	20.26			8.78	6.51	49.57
1955	39.73	37.65	21.91	22.56			13.08	7.18	50.01
1956	44.31	41.91	25.91	26.69			19.06	7.87	51.18
1957	48.93	44.63	29.35	33.09			23.66	15.56	59.20
1958	52.84	45.91	33.87	36.50			28.05	20.84	67.74
1959	56.99	49.36	39.52	41.87			31.81	31.54	72.37
1960	61.19	51.55	42.90	44.34			37.69	44.13	76.75
1961	65.74	55.55	47.86	42.47			44.32	55.88	81.59
1962	70.64	65.50	54.69	45.87			53.22	208.30	86.56
1963	75.30	68.35	60.52	51.11		7.67	61.24	566.77	85.14
1964	79.27	73.66	67.19	62.45	27.24	26.97	79.95	294.31	86.15
1965	85.56	82.04	74.13	73.63	53.51	33.48	85.11	379.02	89.34
1966	92.87	89.68	88.77	89.19	87.21	82.69	89.40	89.74	93.38
1967	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
1968	106.72	107.81	114.46	109.57	122.57	111.59	119.77	111.90	106.87
1969	114.98	122.53	132.62	127.27	149.88	126.50	138.72	133.43	115.53
1970	123.00	130.06	144.63	142.68	186.58	137.60	155.48	150.81	125.46
1971	131.23	133.43	158.57	153.57	227.16	145.32	156.70	146.94	125.01
1972	141.43	144.57	192.69	182.99	285.79	151.59	177.54	519.74	87.23
1973	152.57	161.62	233.58	230.53	383.45	157.58	195.60	1 191.30	72.59
1974	165.70	181.43	291.20	267.39	472.23	154.43	201.95	1 672.25	82.24
1975	179.11	202.19	348.00	309.96	560.71	150.69	228.09	2 333.13	99.30
1976	190.18	217.58	369.09	331.18	647.02	142.20	256.18	2 923.57	119.24
1977	200.12	237.15	396.45	350.62	751.40	137.45	270.54	3 210.50	134.83
1978	208.73	258.96	446.62	410.28	903.84	134.43	294.20	3 263.01	180.27
1979	215.54	270.35	509.35	470.30	1043.97	127.37	314.89	3 537.86	200.67

TABLE B.2: LASPEYRES VOLUME INDICES OF OUTPUT

Year	Local, Directory, Miscellaneous	Toll	Monopoly Toll	Competitive Toll	Message Toll	Other Toll	Gross Production
1952	31.42	21.25	23.86	4.73	24.96	3.69	27.69
1953	34.20	23.13	25.75	6.51	26.94	5.08	30.14
1954	37.19	25.15	27.82	8.24	29.10	6.44	32.77
1955	40.58	29.14	31.80	12.25	33.27	9.57	36.38
1956	44.88	33.38	35.83	17.85	37.48	13.94	40.66
1957	49.77	36.83	39.14	22.19	40.95	17.33	45.02
1958	54.06	39.05	41.05	26.32	42.95	20.56	48.55
1959	58.25	42.86	44.90	29.90	46.97	23.36	52.60
1960	62.47	45.57	47.16	35.47	49.34	27.71	56.27
1961	67.04	48.93	50.06	41.75	52.37	32.61	60.39
1962	71.94	57.16	58.12	51.06	60.80	39.88	66.52
1963	76.10	61.53	61.56	61.40	64.40	47.96	70.76
1964	79.84	70.27	69.02	78.20	70.95	67.04	76.32
1965	85.87	79.27	78.53	83.97	79.69	77.30	83.45
1966	92.91	89.34	89.40	89.00	89.50	88.61	91.60
1967	100.00	100.00	100.00	100.00	100.00	100.00	100.00
1968	106.74	110.76	109.43	119.23	108.82	119.96	108.21
1969	115.03	127.25	125.56	137.95	124.44	140.56	119.51
1970	123.20	138.67	136.20	154.38	133.87	161.43	128.88
1971	130.72	145.25	143.56	155.95	139.70	171.54	136.05
1972	136.98	164.43	162.25	178.27	156.54	201.81	147.05
1973	146.01	191.81	190.52	199.96	181.61	240.13	162.82
1974	158.86	218.56	220.08	208.90	208.43	266.55	180.77
1975	172.57	249.32	251.19	237.49	236.89	308.24	200.74
1976	184.37	270.72	271.29	267.14	253.93	350.30	216.06
1977	194.77	293.56	295.35	282.17	274.29	384.89	231.02
1978	206.39	328.53	332.32	304.43	305.92	435.65	251.22
1979	214.32	357.80	362.94	325.14	331.48	482.51	266.98

TABLE B.3: TÖRNOVIST VOLUME INDICES OF OUTPUT

Year	Local, Dir., Misc.	Toll	Monopoly Toll	Competitive Toll	Gross Production
1952	31.51	21.71	24.32	4.76	27.81
1953	34.29	23.61	26.23	6.55	30.25
1954	37.24	25.65	28.32	8.30	32.86
1955	40.63	29.79	32.47	12.33	36.57
1956	44.90	34.10	36.57	17.96	40.87
1957	49.76	37.54	39.86	22.33	45.20
1958	54.00	39.77	41.79	26.49	48.67
1959	58.19	43.59	45.65	30.09	52.73
1960	62.41	46.30	47.94	35.70	56.38
1961	67.00	49.74	50.96	42.01	60.54
1962	71.92	58.14	59.20	51.48	66.79
1963	76.10	62.52	62.70	61.56	71.05
1964	79.84	70.30	69.05	78.34	76.29
1965	85.87	79.31	78.55	84.12	83.44
1966	92.91	89.34	89.40	89.00	91.60
1967	100.00	100.00	100.00	100.00	100.00
1968	106.74	110.77	109.43	119.23	108.21
1969	115.02	127.27	125.58	137.96	119.50
1970	123.20	138.60	136.11	154.39	128.84
1971	130.79	144.95	143.18	155.95	135.97
1972	141.72	163.60	161.23	178.46	149.82
1973	151.06	190.10	188.45	200.22	165.60
1974	164.35	216.03	217.05	209.09	183.63
1975	178.52	245.95	247.17	237.63	203.68
1976	190.72	267.05	266.86	267.24	219.20
1977	201.46	289.44	290.49	282.25	234.26
1978	213.39	322.86	325.86	304.78	254.06
1979	221.54	349.69	353.81	325.65	269.00

TABLE B.4: OUTPUT PRICE INDICES BY CATEGORY

YEAR	LOCAL	INTRA-BELL	TRANS-CANADA	US, OVERSEAS	WATS	TWX	PRIVATE LINE	MISC. OTHER TOLL	MISC. SERVICES
1952	93.04	106.78	109.94	95.11			98.28	98.28	75.10
1953	93.91	106.74	112.99	95.08			100.79	100.79	75.55
1954	93.96	106.80	114.90	95.12			102.39	102.39	75.44
1955	93.96	106.80	114.91	95.13			102.39	102.39	78.41
1956	93.95	106.79	114.90	94.48			102.38	102.38	78.14
1957	93.78	106.59	114.68	91.92			102.19	102.19	80.12
1958	94.31	107.73	114.60	91.85			102.12	102.12	82.29
1959	100.40	113.76	114.09	91.81			102.08	102.08	88.41
1960	100.35	113.71	113.09	100.79			102.03	102.03	91.37
1961	100.15	111.98	109.73	102.49			101.82	101.82	91.98
1962	100.11	104.44	106.04	102.46			101.91	101.09	93.23
1963	100.11	104.44	104.22	102.46		104.59	102.02	102.04	100.32
1964	100.06	104.38	103.20	102.40	100.26	104.53	101.96	101.86	100.26
1965	100.09	104.41	102.27	102.43	100.19	103.51	101.67	101.48	100.29
1966	100.06	100.78	100.42	102.40	100.06	101.38	100.06	100.11	100.11
1967	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
1968	99.97	98.75	99.87	99.97	99.59	99.53	99.97	99.86	100.98
1969	100.32	99.24	99.67	100.49	99.12	99.58	102.60	101.66	104.28
1970	101.71	111.06	99.76	100.74	99.21	99.67	102.70	101.69	105.08
1971	105.62	113.43	99.67	100.65	98.50	99.58	106.49	103.89	106.10
1972	108.17	115.33	99.22	100.23	97.22	105.40	107.12	103.95	108.42
1973	109.96	117.50	97.99	99.15	98.54	109.34	109.34	105.46	109.77
1974	112.59	119.85	98.22	99.38	102.09	110.32	113.52	108.85	112.00
1975	118.19	122.70	104.16	105.52	105.51	113.79	119.83	113.75	116.86
1976	125.14	128.22	112.07	112.48	108.57	127.06	131.31	121.62	123.63
1977	133.39	133.18	112.39	114.96	110.86	127.78	140.84	127.22	129.91
1978	146.21	144.07	112.92	115.26	117.24	134.68	162.01	140.44	140.23
1979	154.66	153.94	112.82	114.18	123.53	144.29	182.45	152.95	147.59

TABLE B.5: PRICE INDICES OF OUTPUT AGGREGATES

	Törnqvist					Paasche
Year	Local, Directory, Miscellaneous	Toll	Monopoly Toll	Competitive Toll	Gross Production	
1952	90.98	103.11	103.57	97.66	94.82	95.23
1953	91.78	103.27	103.66	100.16	95.43	95.79
1954	91.82	103.46	103.78	101.74	95.52	95.76
1955	92.27	103.41	103.74	101.75	95.82	96.31
1956	92.43	103.32	103.65	101.74	95.90	96.41
1957	92.46	102.83	103.13	101.55	95.78	96.15
1958	93.18	103.59	103.97	101.47	96.51	96.74
1959	99.28	107.61	108.41	101.43	102.00	102.25
1960	99.54	108.79	109.73	101.38	102.53	102.74
1961	99.39	107.59	108.41	101.18	102.07	102.32
1962	99.47	102.28	102.41	101.07	100.47	100.88
1963	100.13	102.21	102.24	101.77	100.90	101.32
1964	100.08	103.55	103.83	101.84	101.31	101.27
1965	100.10	103.42	103.74	101.52	101.28	101.27
1966	100.06	100.88	100.99	100.13	100.36	100.35
1967	100.00	100.00	100.00	100.00	100.00	100.00
1968	100.05	99.23	99.11	99.94	99.75	99.75
1969	100.66	99.92	99.49	102.43	100.38	100.37
1970	102.00	106.61	107.34	102.52	103.82	103.80
1971	105.60	108.37	108.79	106.09	106.70	106.63
1972	104.56	109.30	109.75	106.85	106.42	108.42
1973	106.28	110.51	110.79	109.05	107.92	109.77
1974	108.81	112.59	112.56	113.03	110.25	112.00
1975	114.19	116.93	116.64	119.13	115.17	116.86
1976	120.89	123.63	122.64	130.42	121.86	123.63
1977	128.77	127.94	126.23	139.39	128.11	129.91
1978	141.00	136.56	133.11	159.57	138.66	140.23
1979	149.10	144.11	138.89	179.14	146.48	147.59

TABLE B.6: LABOUR VOLUME INDICES

YEAR	UNADJUSTED HOURS WORKED	ADJUSTED HOURS WORKED	
		LASPEYRES	TÖRNQVIST
1952	85.54	79.40	81.63
1953	86.57	81.41	83.02
1954	91.51	85.17	87.29
1955	99.11	91.71	94.50
1956	106.47	98.38	102.46
1957	110.60	102.15	106.87
1958	108.26	101.80	105.32
1959	101.72	99.91	99.97
1960	97.30	96.50	96.23
1961	91.54	92.69	91.16
1962	91.19	92.40	90.77
1963	94.01	94.59	93.06
1964	95.60	96.19	94.57
1965	98.09	98.62	97.59
1966	103.06	101.57	102.39
1967	100.00	100.00	100.00
1968	96.43	98.07	98.41
1969	98.15	100.03	100.05
1970	99.21	102.22	102.66
1971	97.50	101.45	101.73
1972	97.43	101.71	101.80
1973	102.24	106.71	106.77
1974	108.82	112.86	112.74
1975	108.42	113.38	113.09
1976	113.58	119.01	118.57
1977	117.67	123.36	123.01
1978	125.81	132.98	132.54
1979	129.20	137.05	137.07

TABLE B.7: LABOUR PRICE INDICES

YEAR	UNADJUSTED HOURS WORKED	ADJUSTED HOURS WORKED	
		PAASCHE	TÖRNQVIST
1952	45.37	48.88	47.54
1953	49.38	52.51	51.49
1954	50.97	54.77	53.44
1955	52.84	57.10	55.41
1956	54.01	58.46	56.13
1957	56.34	61.00	58.31
1958	60.46	64.30	62.15
1959	66.22	67.42	67.38
1960	71.24	71.83	72.03
1961	76.92	75.97	77.23
1962	80.34	79.29	80.71
1963	82.36	81.85	83.20
1964	84.45	83.92	85.36
1965	86.98	86.51	87.43
1966	91.45	92.78	92.04
1967	100.00	100.00	100.00
1968	110.36	108.52	108.14
1969	119.84	117.59	117.04
1970	133.85	129.91	129.35
1971	144.63	139.00	138.61
1972	162.87	156.02	155.87
1973	175.57	168.21	168.11
1974	196.32	189.29	189.48
1975	235.87	225.56	226.14
1976	266.14	254.00	254.95
1977	295.74	282.11	282.92
1978	312.92	296.05	297.03
1979	360.74	340.07	340.03

TABLE B.8: DISAGGREGATED LABOUR PRICES

(1952-1967)

Year	Telephone Operators	Plant Craftsmen	Clerical	Other Employees
1952	1.28	1.85	1.47	1.85
1953	1.34	1.94	1.53	2.12
1954	1.41	1.98	1.57	2.21
1955	1.47	2.00	1.59	2.35
1956	1.50	2.01	1.62	2.36
1957	1.59	2.09	1.71	2.40
1958	1.69	2.21	1.82	2.58
1959	1.77	2.48	1.91	2.83
1960	1.84	2.52	1.99	3.18
1961	1.91	2.65	2.06	3.54
1962	1.94	2.75	2.11	3.78
1963	1.98	2.83	2.16	3.92
1964	2.03	2.90	2.22	4.02
1965	2.03	3.01	2.31	4.10
1966	2.11	3.14	2.38	4.39
1967	2.27	3.41	2.57	4.79

TABLE B.9: DISAGGREGATED LABOUR PRICES
(1967-1979)

Occupation Groups and Years of Service	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
Telephone Operator													
- 1	$ 1.997	$ 2.148	$ 2.308	$ 2.577	$ 3.088	$ 3.078	$ 3.240	$ 3.922	$ 4.672	$ 5.141	$ 5.581	$ 5.613	$ 5.729
1 - 2	2.099	2.354	2.583	2.867	3.434	3.716	3.904	4.319	5.312	5.903	6.325	6.497	6.512
3 - 5	2.373	2.551	2.782	3.161	3.772	4.078	4.495	5.143	6.079	6.464	7.248	7.594	7.615
6+	2.569	2.758	2.963	3.269	3.911	4.211	4.633	5.239	6.413	6.848	7.641	7.804	7.862
Plant Craft													
- 1	2.504	2.613	2.765	3.107	3.379	3.741	3.935	4.599	5.173	5.130	6.174	6.864	8.662
1 - 2	2.747	3.093	3.354	3.572	3.972	4.362	4.769	5.476	6.069	6.370	7.238	8.221	10.444
3 - 5	3.295	3.470	3.901	4.604	5.060	5.228	5.800	6.835	7.563	7.447	8.839	10.064	12.751
6 - 8	3.665	3.945	4.309	4.816	5.256	5.742	6.237	7.266	8.015	7.969	9.517	10.530	13.519
9+	3.842	4.094	4.439	4.968	5.492	6.021	6.536	7.569	8.310	8.315	9.927	10.980	13.952
Clerical													
- 1	2.128	2.272	2.461	2.718	2.850	3.086	3.367	3.962	4.400	5.322	5.872	6.700	6.842
1 - 2	2.316	2.601	2.804	3.053	3.316	3.583	3.860	4.594	4.918	6.235	6.685	7.867	7.853
3 - 5	2.613	2.813	3.122	3.551	3.899	4.226	4.663	5.453	5.839	6.987	7.661	8.975	9.153
6+	3.040	3.246	3.482	3.822	4.196	4.583	5.021	5.880	6.369	7.629	8.284	9.437	9.568
Other Non-Management													
- 1	2.722	2.934	3.311	3.597	3.799	4.241	4.553	5.180	5.545	6.351	7.145	8.135	8.953
1 - 2	3.098	3.332	3.505	3.958	4.314	4.628	5.000	5.904	6.358	7.582	8.190	9.470	10.360
3 - 5	3.294	3.554	3.889	4.318	4.582	4.954	5.515	6.378	6.943	8.082	9.074	10.378	11.376
6+	3.672	3.928	4.181	4.649	4.862	5.295	5.793	6.812	7.392	8.416	9.509	10.814	11.809
Foreman & Supervisor													
- 5	3.561	3.966	4.430	5.069	5.406	6.035	6.684	7.170	9.455	9.859	10.626	11.433	12.191
5 - 9	3.888	4.204	4.534	5.322	6.012	6.753	7.401	8.336	10.608	11.341	12.406	13.291	14.117
10 - 14	4.557	4.918	5.324	5.990	6.453	7.097	7.804	8.734	11.248	12.290	13.513	14.402	15.255
15+	5.264	5.799	6.122	6.877	7.630	8.474	9.181	10.318	12.739	14.032	15.283	16.153	16.732
Other Management													
- 5	4.391	4.675	5.006	5.631	6.031	6.672	7.393	8.234	10.450	11.486	12.415	13.365	14.138
5 - 9	5.498	5,973	6.261	6.793	7.446	8.248	8.941	9.692	12.288	13.925	15.320	16.023	17.051
10 - 14	5.822	6.600	6.950	7.817	8.625	9.339	10.067	10.813	13.237	14.678	16.167	16.818	17.946
15 - 19	6.317	6.460	7.138	8.036	8.910	9.913	11.009	12.209	14.916	16.514	17.690	18.482	19.435
20+	7.139	7.531	8.049	9.004	9.921	10.945	11.848	12.591	15.076	16.512	18.076	18.972	20.335
Part Time	3.525	3.959	4.230	4.700	5.468	5.703	5.988	7.623	9.103	10.049	10.926	11.440	11.201
Total	3.384	3.721	4.014	4.537	5.066	5.530	6.009	6.811	7.940	8.734	9.748	10.742	11.824

TABLE B.10: INDICES OF NET CAPITAL STOCK

YEAR	UNADJUSTED	ADJUSTED (TÖRNQVIST)		
		RESIDUAL RATE OF RETURN WEIGHTS		USER COST WEIGHTS
	WITH PUC		WITHOUT PUC	
1952	27.28	26.79	26.85	26.48
1953	30.05	29.54	29.61	29.21
1954	32.85	32.34	32.43	32.06
1955	36.76	36.12	36.07	35.74
1956	41.12	40.65	40.65	40.30
1957	46.02	45.62	45.63	45.29
1958	51.35	50.98	50.97	50.62
1959	56.68	56.33	56.44	56.03
1960	62.19	61.85	62.11	61.62
1961	67.34	67.10	67.53	67.02
1962	72.37	72.19	72.65	72.19
1963	77.82	77.84	78.35	77.96
1964	83.11	83.31	83.81	83.53
1965	88.33	88.62	89.10	88.89
1966	94.07	94.37	94.79	94.68
1967	100.00	100.00	100.00	100.00
1968	105.74	105.78	105.68	105.68
1969	111.93	112.03	111.91	111.89
1970	117.91	117.82	117.48	117.49
1971	124.35	124.02	123.32	123.33
1972	131.28	131.04	130.32	130.30
1973	137.40	137.24	136.60	136.61
1974	144.44	143.95	142.99	142.98
1975	153.02	152.22	151.53	151.26
1976	161.41	159.93	159.57	159.39
1977	169.56	168.51	168.36	168.28
1978	174.97	174.44	174.62	174.67
1979	179.35	178.70	178.86	179.07

TABLE B.11: TÖRNQVIST INDICES OF NET CAPITAL PRICE

	RESIDUAL RATE OF RETURN		USER COST
	PLANT UNDER CONSTRUCTION		
YEAR	INCLUDED	EXCLUDED	
1952	73.09	72.95	64.26
1953	72.74	72.56	64.85
1954	70.75	70.55	59.72
1955	69.45	69.55	58.02
1956	67.94	67.93	60.20
1957	70.21	70.19	70.20
1958	69.49	69.51	70.60
1959	80.49	80.34	77.65
1960	81.69	81.35	79.76
1961	83.72	83.19	78.69
1962	86.70	86.15	80.86
1963	86.31	85.74	81.58
1964	89.79	89.25	81.76
1965	93.19	92.69	81.38
1966	94.47	94.04	92.31
1967	100.00	100.00	100.00
1968	102.72	102.81	109.59
1969	104.87	104.97	115.01
1970	112.75	113.08	124.73
1971	113.18	113.82	118.87
1972	117.67	118.32	125.30
1973	125.75	126.34	139.89
1974	131.20	132.08	163.77
1975	143.67	144.33	178.49
1976	150.88	151.22	183.94
1977	154.36	154.50	197.97
1978	179.04	178.85	218.16
1979	189.98	189.81	247.62

TABLE B.12: MATERIAL VOLUME AND PRICE INDICES

	VOLUME INDEX		PRICE INDEX	
YEAR	LASPEYRES	TÖRNQVIST	PAASCHE	TÖRNQVIST
1952	38.87	38.87	74.20	74.20
1953	41.75	41.75	74.00	74.00
1954	46.73	46.74	75.20	75.20
1955	53.49	53.49	75.70	75.70
1956	62.69	62.69	78.50	78.50
1957	63.19	63.19	80.10	80.10
1958	69.47	69.47	81.30	81.30
1959	73.12	73.12	82.90	82.90
1960	76.43	76.43	83.90	83.90
1961	79.69	79.69	84.30	84.30
1962	85.42	85.42	85.40	85.40
1963	89.92	89.92	87.10	87.10
1964	90.17	90.17	89.20	89.20
1965	98.35	98.35	92.10	92.10
1966	102.29	102.29	96.20	96.20
1967	100.00	100.00	100.00	100.00
1968	104.63	104.63	103.30	103.30
1969	124.66	124.66	107.80	107.80
1970	124.63	124.63	111.85	111.85
1971	148.03	147.96	116.04	116.09
1972	151.54	151.45	119.92	119.99
1973	160.47	160.28	125.66	125.81
1974	169.60	169.23	138.27	138.57
1975	168.01	167.60	152.44	152.81
1976	181.12	180.75	165.92	166.27
1977	202.90	202.41	179.85	180.29
1978	222.42	221.54	192.94	193.90
1979	224.08	222.22	209.33	211.09

TABLE B.13: DEPRECIATION RATES (%)

YEAR	BUILDINGS	COE	STATION	OUTSIDE PLANT	GENERAL EQUIP.
1952	2.4	4.7	5.7	3.6	7.7
1953	2.4	4.8	5.8	3.6	7.6
1954	2.4	4.5	5.9	3.6	7.7
1955	2.4	4.4	5.9	3.4	7.7
1956	2.3	4.4	6.1	3.4	7.8
1957	2.3	4.4	7.5	3.4	8.0
1958	2.3	4.5	7.5	3.4	8.7
1959	2.4	4.7	7.9	3.4	8.9
1960	2.5	4.6	7.9	3.4	8.9
1961	2.5	4.3	8.2	3.5	9.1
1962	2.5	4.3	8.2	3.6	9.1
1963	2.5	4.4	8.8	3.6	9.1
1964	2.5	4.4	8.9	3.5	9.1
1965	2.6	4.5	9.0	3.5	9.4
1966	2.6	4.6	9.1	3.5	9.3
1967	2.6	4.6	9.2	3.5	9.3
1968	2.7	4.6	9.2	3.4	9.2
1969	2.7	5.0	9.2	3.4	9.2
1970	2.7	5.0	9.1	3.3	9.2
1971	2.8	4.8	9.1	3.4	9.6
1972	3.2	5.0	9.6	3.5	10.0
1973	3.3	5.2	9.8	3.5	13.0
1974	3.4	5.0	10.6	3.5	13.8
1975	3.5	4.9	11.7	3.7	13.1
1976	3.7	5.1	11.6	3.7	11.1
1977	3.3	5.1	11.6	3.7	10.7
1978	3.3	5.0	11.4	3.8	11.3
1979	3.3	5.1	11.2	3.9	11.1

COMMENT on

- Fuss, Mel (University of Toronto), "Survey of Recent Results in Analysis of Production Conditions in Telecommunications",
- Kiss, Frank, S. Karabadjian, and B. Lefebvre (Bell Canada). "Economies of Scale and Scope in Bell Canada: Some Econometric Evidence",
- Christensen, Laurits, Dianne Cummings, Philip E. Schoech (University of Wisconsin). "Econometric Estimation of Scale Economies in Telecommunications",

(papers presented to the Production Analysis session of the Conference "Telecommunications in Canada", Montreal, March 4, 1981)

by Erwin A.J. Dreessen*

* Economist, British Columbia Telephone Company, currently on loan to the TransCanada Telephone System. Views expressed are his own and do not necessarily represent those of B.C. Tel or TCTS.

1981 03 15
Rev. 1981 06 15

My comments are directed at the papers by Christensen et. al. (CEA) and Kiss et. al. (KEA) and underscore Fuss' conclusion that estimates of scale are not invariant to the specification of technological change. All three papers are valuable contributions to the literature, yet much work remains to be done. I suggest some directions.

Two sets of variations in technology specification can be distinguished, one relating to different choices for the technology variable, the second to whether or not the technology variable(s) is (are) allowed to interact with other variables, including whether technology is allowed to be output- or factor-augmenting.

The favoured measure of technological change in CEA is a distributed lag of R&D expenditures by AT&T. Its advantage is that R&D expenditures can be expected to have a pervasive effect on the operating companies and are therefore relatively easy to model in a general way. The disadvantages, as Terleckyj has observed[1], are that they measure only a portion of the total investment in innovation, that the rate of depreciation of R&D capital is uncertain, and that the conceptual and expected statistical links between such expenditures and indices of input and output are unclear.

CEA report on alternative, including "direct" measures of technological change. Some of these result in appreciably higher estimates of scale economies, especially when allowance is made for serial correlation.

[1] Nestor E. Terleckyj, "What Do R&D Numbers Tell Us About Technological Change?", A.E.R., vol. 70, no. 2 (May 1980), 55-61.

However, these results cannot be fully evaluated without knowing how the preferred R&D measure differs from the alternatives. KEA do provide that information. The level and pattern of their preferred measure is very close to that of two alternatives, while a fourth measure exhibits substantially slower growth; employment of the latter measure results in a substantially higher estimate of scale elasticity.

The direct measures considered in both papers, as well as in most studies surveyed by Fuss, reflect primarily changes in access to DDD and/or changes in switching technology. Such indicators are less ambiguous than an R&D measure. However, they reflect only part of the technological changes occurring in the industry(2). Their very specificity imposes a burden on the researcher to cast a wide eye on technological developments.

It would appear that direct measures of technological change can more plausibly be put into a restrictive framework (a practical requirement due to data limitations). Additional specific information regarding technology could be brought to bear on multiple-output models. Yet CEA's paper is the one that experiments most widely with second-order and interaction terms. It also has one model with factor-augmenting R&D, but none with output-augmenting technology. None of their experiments lend credence to

(2) Fuss mentions the omission of small-scale technological change and improvements in transmission technology. Other omissions that immediately come to mind are replacements of manual operator functions by machines or by automated operator functions, and the application of information retrieval systems in areas such as customer services and directory assistance. As well, of course, there is the gradual penetration of digital transmission systems since the mid-60s and of digital switches after 1975.

the hypothesis of constant or decreasing returns to scale, but nor are the resulting estimates stable. In particular, models (7), (8), (11) and (12), which include second-order terms in both technology and output, estimate the scale elasticity to be considerably higher.

KEA, in their two-output models, proceed from a very general form to truncations which do away with some output terms, including an output-technology interaction term. As a result, the estimates of scale economies increase somewhat, and output-specific scale estimates become meaningful.

Finally, Fuss reports on Fuss and Waverman (1980) who, assuming partial profit maximization and output-augmenting technology, find e to be 0.94 under the Box-Cox transformation of variables.

The lesson is clear: Scale elasticity estimates are not robust with respect to alternative specifications of technology; in particular, the estimates are sensitive to the presence or absence of interaction terms between output and technology.

The set of estimates presented in the papers leaves areas to be explored. The discussion would benefit by more explicit reporting on multicollinearity[3].

More engineering information on technological changes and what they affect would help identify reasonable restrictions -- strict adherence to likelihood ratio tests can lead one astray. Single-output models must not

[3] If multicollinearity is serious then Vinod's suggestions for modified ridge regression deserve to be taken up. Cfr. Hrishikesh D. Vinod, "Application of New Ridge Regression Methods to a Study of Bell System Scale Economies", *Journal of the American Statistical Association*, (December 1976), Vol. 71, No. 356, pp. 835-841.

be neglected, if only because they provide a vehicle for learning to overcome certain technical problems, but in the final analysis they can only serve to underscore and build confidence in multiple-output models. Policy relevance is to be found in the latter. Permissible output aggregation needs more attention; local output may have to be unbundled in connection and terminal charges.

It is tempting to despair of disentangling technological change and (dis)economies internal to a firm. But evidence on cost complementarity and scale and scope economies would loose all policy relevance if technological change were completely embedded.

COMMENTS ON THE PAPERS IN THE PRODUCTION ANALYSIS SESSION

J.B. SMITH
CONCORDIA UNIVERSITY

In my comments I will examine three issues which link many of the papers in this section and which I feel have straightforward policy implications. In addition, I will examine two results which I am hard pressed to explain and which I feel should be the subject of greater attention. First however I would like to state that all of the papers are very interesting and well-prepared.

The first issue has to do with the existence of a binding A-J type rate-of-return constraint in telecommunications. None of the papers presented here incorporate it and the history of attempts (at least for Bell Canada data) is one of failure. I don't think that Bell Canada is bound (in the classic sense) by an A-J constraint. As well, however, I don't think that elaborate econometric models are necessary to pull out this result. It is sufficient to examine the factor cost shares of Bell Canada and to note that for long periods of time they (in particular, materials) are effectively constant. Simple regression can be used to establish the statistical significance of this result. It then remains to note that within the A-J framework constant factor shares are ruled out and thereby conclude that regulation must be exerting its influence only through output prices. Whether or not this result holds true for the U.S. is yet to be determined.

The second issue concerns the way in which technical change and scale economies are being traded-off in the estimation process. It would appear that the Canadian data support two sets of results: the first where technical gains are high and scale is low (and often trended) and the second where scale is roughly constant and significantly greater than one with technical progress being less important. While Christensen's results seem to support the second result for U.S. data, it would appear that no hard and fast conclusion can be drawn at this time. (The results of Denny *et al*. don't clarify this issue since they assume constant returns to scale.) One statistical way out of this impass might be to employ nested and/or non-nested techniques to compare the models. In addition, should the U.S. and other Canadian data become public, a method might be found whereby the samples could be pooled. Until such time as one or both of these suggestions is implemented, it would appear that strong public policy conclusions are not possible.

The third issue deals with the amount of information which researchers are willing to introduce into the estimation process. As noted by Professor Fuss, some studies have included the assumption that one or more service outputs of Bell Canada are supplied at a rate which maximizes profits. Others (including the Christensen and Kiss studies) have not included this assumption. The Kiss study stands out to the extent that it is one of the few multi-output studies using

Bell Canada data which does not assume profit maximization. The Arguments in favour of including the profit maximization assumption revolve around the fact that added information, if accurate, will improve the efficiency of the parameter estimates. This is particularly important when the sample is a time series of highly trended variables and when many parameters must be estimated. Mr. Kiss and others have argued that the elasticity restrictions which are required to make marginal revenues positive are not reasonable for message toll service output of Bell Canada. Their multi-output econometric results without this assumption demonstrate the classic effects of multicollinearity. As with issue two, the resolution of this problem rests in increasing the sample. Professor Diewert has suggested some nonparametric tests which might be used to examine the validity of the profit maximization assumption. Finally, it should be noted that some of the models which incorporate the profit maximization assumption simultaneously supply weak support for the assumption when they satisfy the second order conditions for profit maximization. As well, these latter models tend to track and forecast quite accurately.

Turning to the results which I find difficult to understand, the first relates to the Denny finding that there are potentially significant productivity differences between the major Canadian telecommunications carriers. This result emerges from a first pass at the data and I suspect it will be the case than when data differences are ironed-out, the productivity groupings will be much closer.

The second problem relates to the fact that Christensen estimates the output scale elasticity of ATT to be approximately 1.5. Many studies using Bell Canada data estimate scale elasticities to be about the same value. On the face of it, the consistency of the results may appear encouraging. However, this consistency has important implications for comparing ATT and Bell technologies. In particular, it must be recalled that based upon 1975 data, ATT is approximately 17 times larger than Bell. If we write the cost functions for Bell and ATT in the isoelastic form as:

$$C_i = f_i(p_i, t_i) q_i^{2/3} \qquad i = \text{Bell, ATT}$$

where f_i is homogeneous of degree 1 and (weakly) concave in factor prices p_i and t_i is a variable representing the level of technology we can gain some insight into these implications. For example, if factor prices and technology levels are the same and further if f is the same, the implication is that the marginal and average costs of Bell must be approximately 2.6 times larger than those of ATT. If we drop the assumption that factor prices and technology levels are the same and replace it with the reasonable alternative that ATT prices are lower and the technology level higher, we find that the divergence between Bell and ATT unit costs must increase. Thus, if we believe that U.S. and Canadian costs do not differ by the amounts indicated above, we are left with the conclusion that the f function must differ. It would seem important therefore to study more carefully why these differences arise. This would require a coordination of U.S. and Canadian research efforts. Unfortunately, the U.S. data are presently restricted and the implementation of this suggestion would require careful negotiations.

MANAGEMENT APPLICATIONS OF PRODUCTIVITY ANALYSIS

GLOBAL FACTOR PRODUCTIVITY (GFP)
AND EDF'S MANAGEMENT

J.N. REIMERINGER

Electricité de France

INTRODUCTION

La gestion d'une entreprise n'est - et ne peut pas être - indépendante des objectifs qu'elle poursuit.

Or Electricité de France, à sa création au lendemain de la dernière guerre mondiale, a reçu une mission de Service Public : sous le contrôle de l'Etat, mais dotée d'une autonomie de gestion, elle exerce une activité industrielle et commerciale conformément à l'intérêt général. La manière de le faire n'est pas définie dans la loi de nationalisation, qui se contente d'indiquer qu'elle doit "faire face" aux dépenses nécessaires à l'exercice de son autorité.

L'entreprise a donc été amenée peu à peu à fixer elle-même les critères de son action dans le souci de l'intérêt de la collectivité. Ils se définissent à partir de quatre principes essentiels :

- l'égalité de traitement des usagers devant le Service Public : du fait qu'EDF a le monopole du transport et de la plus grande partie de la distribution d'électricité, la conséquence directe en est que, hors cas de force majeure, tout client a le droit d'exiger que l'Etablissement lui livre à tout instant l'énergie électrique qu'il demande. La traduction

...

pratique en est l'utilisation dans le calcul économique d'un coût de défaillance, unique et élevé, qui exprime dans un contexte aléatoire le niveau de sacrifice financier que l'Etablissement supporte pour éviter de délester (1).

- la vente au coût marginal (en développement) : le prix n'est ni un "prix de marché", c'est-à-dire formé selon la loi de l'offre et de la demande, ni un prix "rémunérateur". Dans la limite des contraintes financières, il est calé sur le coût marginal de la fourniture : il permet ainsi d'orienter les choix de la clientèle de manière à éviter le gaspillage.

- la référence au taux d'actualisation du plan : les investissements sont choisis de façon à minimiser le coût total actualisé de gestion courante (entretien, exploitation, combustible...), de défaillance et d'investissement des équipements, le taux d'actualisation choisi n'étant pas un taux interne (c'est-à-dire par exemple le taux de rentabilité maximum compatible avec les possibilités financières de l'Etablissement), mais un taux déterminé par les Pouvoirs Publics, dans le cadre de leur politique de développement national à long terme, qui trouve en France sa concrétisation dans le Plan.

- enfin, la nécessité pour l'entreprise d'être de plus en plus performante : or, pour une entreprise qui fixe ses prix en dehors de toute considération de rentabilité financière, il ne peut s'agir de maximiser son profit. Aussi a t-il été nécessaire d'élaborer un autre critère, qui permette à la fois à l'entreprise de mesurer globalement les progrès que ses choix ont entraînés sur sa performance, et à la collectivité de juger si cette dernière s'améliore. A ces deux préoccupations répond un outil : la mesure de la productivité globale des facteurs.

- 0 -

- 0 - - 0 -

...

(1) C'est-à-dire d'effectuer des coupures pendant un certain temps (généralement très court).

I - EXPOSE SUCCINCT DE LA METHODE

On peut définir la performance d'une entreprise, par rapport à la collectivité, comme sa capacité à produire plus en consommant moins. Sa mesure implique que l'on comparera dans le temps plusieurs états de l'entreprise, et qu'on raisonnera alors en termes de productivité, c'est-à-dire de rapport quantités de produits/quantités de facteurs de production. Mais ce rapport, pour être significatif, est loin d'être simple à établir.

1.1. - Productivité du travail et productivité globale

Le terme de productivité est devenu, dans le langage courant, synonyme de productivité du travail, elle-même définie comme le rapport entre la production ou la valeur ajoutée (1) et la quantité de travail. Or l'amélioration de la productivité du travail ne signifie pas nécessairement que l'entreprise est devenue plus "performante", au sens qui nous intéresse ici. Imaginons par exemple un atelier qui, avec 10 ouvriers, produit journellement 100 unités d'un bien donné, disons des chaises. Un changement de machines permet de produire 200 chaises avec la même quantité de matières par chaise. La productivité du travail a certes doublé, mais le gain collectif n'est pas double. Les machines ont coûté elles-mêmes un certain nombre d'heures de travail - qu'on appelera indirect - et une certaine quantité de matières premières. Si l'on veut avoir une idée "globale" de la productivité de l'atelier, il faut que figurent au dénominateur tous les facteurs impliqués dans le processus de production. On peut alors définir une mesure de la performance de l'entreprise au cours d'une période donnée comme la variation sur cette période de la productivité globale.

1.2. - Principe de l'évaluation de la productivité globale

Le rapport production/somme des facteurs semble à première vue simple à calculer. Ne suffit-il pas de faire le rapport produits/charges en partant des montants qui figurent au crédit et au débit du compte d'exploitation générale ? La réponse est évidemment non. Les chiffres comptables sont en effet des valeurs, c'est-à-dire des quantités de biens multipliées par des prix. Si, tout en produisant les mêmes quantités de biens avec les mêmes facteurs, l'entreprise double ses prix sans que ceux des facteurs ne changent, le rapport produits/charges est multiplié par deux alors que la performance de l'entreprise reste la même.

La productivité doit être évaluée sur la base d'un rapport de quantités (on dit encore de volumes), et non de valeurs.

Mais s'il est facile d'additionner des valeurs (qui toutes s'expriment en francs), il est beaucoup plus ardu de faire la somme de quantités de biens aussi disparates que des heures de travail, des tonnes de pétrole, des machines et des kWh. Il faut un jeu de coefficients d'équivalence qui soit cohérent avec le but recherché : la plus grande satisfaction au moindre coût.

...

(1) valeur ajoutée = valeur des produits - coûts des matières premières (dont combustibles)

Le choix de ce jeu de coefficients ne se pose pas de la même façon pour les produits et pour les charges.

Les produits

E.D.F. a une maîtrise relative de sa tarification, notamment en structure, et la définit de manière à ce que le prix du kWh desservi à tout moment et à tout niveau de tension reflète au mieux son coût marginal (1).

Dès lors, d'une année sur l'autre, toute croissance ou modification structurelle de la consommation trouve sa contrepartie dans les recettes, proportionnellement à l'effort nécessaire à l'Etablissement pour réajuster sa production à la nouvelle demande. Les prix représentent des coefficients d'équivalence significatifs.

Néanmoins, un problème demeure : celui de la qualité de service. Il est en effet impossible de desservir les clients à tension parfaitement constante, et les aléas de production et distribution conduisent dans certains cas, heureusement très rares, à "délester", c'est-à-dire à priver d'électricité pendant un certain temps, généralement court, une partie de la clientèle.

Or, paradoxalement, lorsque l'entreprise est obligée de couper des clients, ce "résultat" n'a pas de prix, et n'est donc pas considéré comme produit.

Cependant, au niveau du système de production, la qualité de service est prise en compte par l'intermédiaire du kWh défaillant, dont le coût (normatif) sert de base au choix des investissements. Par cohérence, il serait donc logique de l'introduire comme nouveau produit.

Cela ne poserait pas de problème pour le calcul prévisionnel : pour un parc donné et une certaine prévision de demande, on sait calculer en espérance le coût de la défaillance. Il suffit de reconstruire un compte d'exploitation avec une recette négative correspondant à ce coût, et on peut alors mesurer exactement la performance de l'entreprise.

Mais pour qu'un calcul prévisionnel ait un intérêt, il faut pouvoir le comparer au calcul ex-post, après réalisation. Or, on ne réalise pas de la défaillance, on fait (éventuellement) des coupures, lesquelles dépendent des pannes, des importations, des conditions climatiques, de la pluviosité... La correspondance entre les deux est pour le moins aléatoire, la différence provenant de phénomènes "externes", indépendants de la gestion de l'établissement.

...

(1) Il ne peut pas lui être rigoureusement égal pour divers raisons :

La première d'ordre technique, dans la mesure où il est impossible de comptabiliser les kWh consommés à tout moment, mais seulement par tranches horaires, la deuxième d'ordre économique, les prix reflétant les coûts marginaux à long terme (car devant orienter les consommations), la troisième d'ordre financier, l'équilibre budgétaire devant être globalement réalisé.

Force est donc de renoncer, au niveau de la production, à valoriser les kWh coupés (ce serait peut-être possible en distribution, où la loi des grands nombres joue). Ne pas en tenir compte n'est pas satisfaisant, car cela aboutit à une situation particulièrement paradoxale : pire est la qualité du service, meilleure est la performance. En effet, lorsque le parc est sous-ajusté (1), l'influence sur les produits est faible (elle est égale au "manque à gagner" du fait des kWh coupés, s'il y en a), alors que les charges sont inférieures à ce qu'elles seraient si le parc était parfaitement ajusté. Dès lors, la performance de l'entreprise d'une année sur l'autre est d'autant plus élevée que le parc de la seconde année est plus sous-ajusté par rapport à la première, ce qui est pour le moins paradoxal.

Une solution consisterait à éliminer purement et simplement l'aspect qualité de service. Il suffit pour cela de ramener chaque année à une espérance de défaillance équivalente, en réajustant fictivement le parc. Si, par suite d'une surestimation de la demande ou d'un arbitrage dynamique, on a un parc sur-ajusté, on retranche le "surplus" de puissance.

Cette solution présente deux avantages : tout d'abord elle est simple au point de vue calcul, ensuite elle conduit à une valorisation nulle quand on est à l'optimum à long terme. Or, par définition, l'optimum résulte d'un arbitrage coût-défaillance, lequel se retrouve dans les coûts marginaux à long terme, signal des prix. Quand le parc est simplement ajusté, mais non optimum, on devrait en fait valoriser la marge entre l'espérance de défaillance à parc ajusté et celle à parc optimal. Cependant, on fait alors l'hypothèse implicite, mais non nécessairement exacte, que le surcroît de qualité de service ainsi réalisé est équivalent pour la collectivité à la gêne entraînée par un sous-ajustement de la même ampleur.

La correction de l'aspect qualité de service pose donc encore des problèmes d'ordre méthodologique et pratique difficiles à résoudre. Heureusement le parc est pratiquement toujours ajusté, et la correction envisagée ne serait que de faible importance. Pour toutes ces raisons, le calcul actuel n'en tient pas compte.

<u>Les charges</u>

Pour être en pleine cohérence avec l'objectif poursuivi qui est de mesurer une performance par rapport à la collectivité, il faudrait valoriser les facteurs de production à leurs coûts marginaux. Or, ces derniers ne sont pas connus, et les prix des facteurs s'en écartent pour de nombreuses raisons : rapports de force entre états, rigidités structurelles, intervention tions de l'Etat, politique des entreprises privées en situation de monopoles ou d'oligopoles... D'un autre point de vue, ce sont les prix qui caractérisent l'environnement de l'entreprise, par rapport auquel elle prend ses décisions de gestion. En prenant un système de coefficients d'équivalence autre, on mesurerait une performance différente de l'objectif de gestion de l'entreprise.

...

Il faut néanmoins distinguer deux types de charges :

- les charges d'exploitation : elles correspondent à la gestion de l'outil de production à un instant donné. Cette gestion vise à minimiser le coût de production à court terme, compte tenu de l'environnement économique instantané. Ce sont donc les prix du marché à cet instant là qui dictent la conduite de l'entreprise. Ce sont donc eux qu'il faut utiliser pour mesurer la performance de l'entreprise, telle que précisée ci-dessus.

  On effectue cependant une correction d'une autre nature : la correction d'hydraulicité. En effet, on veut une mesure de l'efficacité de la gestion de l'entreprise. Il faut donc éliminer du calcul les effets de phénomènes externes, sur lesquels l'entreprise n'a pas prise, dont notamment la pluviosité, qui peut faire varier considérablement le productible hydraulique. On se ramène à une hydraulicité "normale".

- les charges de capital : Les charges de capital se décomposent, dans la comptabilité de l'entreprise, en :

  - charges financières : ce sont les intérêts des emprunts que l'entreprise doit contracter pour acquérir les divers biens (immeubles, machines, stocks...) qui constituent le capital ;

  - dotations aux amortissements : elles représentent la dépréciation du capital, du fait de son usure et de l'évolution de la technologie (il devient obsolète).

Or, par rapport au but fixé, qui est de mesurer la performance de l'entreprise, les charges comptables introduisent un biais pour trois raisons principales :

- d'abord parce qu'elles sont exprimées en francs courants. Un capital de 1 000 F ne représente pas la même quantité de biens suivant qu'il a été investi en 1960 ou 1978, à cause de l'érosion monétaire La première correction consistera donc à raisonner en francs constants, c'est-à-dire à réévaluer l'ensemble des actifs ;

- ensuite parce que les charges financières dépendent des taux d'intérêt des divers emprunts que contracte l'entreprise, qui peuvent être très variables. Or nous voulons mesurer la performance d'une entreprise publique, dont une des caractéristiques est sa capacité d'arbitrer conformément à l'intérêt général entre une dépense de capital (et donc un surcroît de peines) dans l'instant, et des économies (et donc un surcroît de satisfaction) dans le futur.

  Il faut par conséquent connaître la préférence *collective* entre une dépense aujourd'hui et une économie demain. Il s'agit de déterminer un taux de "préférence sociale pour le futur", c'est-à-dire le "taux d'actualisation". On rappelle que ce taux (par exemple a) est tel qu'il est indifférent à la collectivité de consommer des biens d'une valeur X aujourd'hui et des biens d'une valeur X (1 + a) (après correction de l'inflation) l'année prochaine.

Ce taux est fixé par le Commissariat Général au Plan (il est actuellement de 9 %). Il constitue une des bases des décisions d'investissement prises par l'Etablissement, et une des conventions essentielles du calcul économique. Par souci de cohérence, on est amené à prendre en compte ce même taux lors de l'évaluation des performances économiques.

- enfin parce que l'amortissement comptable du capital ne représente pas réellement la dépréciation des équipements, au sens du calcul économique. Il y a donc lieu de procéder à certaines corrections, en cohérence avec les objectifs poursuivis.

Les charges de capital ainsi recalculées sont appelées *normatives*, par opposition aux charges inscrites aux comptes, dites effectives.

1.3. - Formulation mathématique (1)

On notera :

q_i^1 : quantité du bien i produit l'année 1

q_i^0 : quantité du bien i produit l'année 0

p_i^0 : prix du bien i l'année 0

f_j^1 : quantité du facteur j consommée l'année 1

f_j^0 : quantité du facteur j consommée l'année 0

p_j^0 : prix du facteur j l'année 0

On peut alors calculer :

. l'indice de croissance de la production $P_{1/P_0} = \dfrac{\sum_i p_i^0 \, q_i^1}{\sum_i p_i^0 \, q_i^0}$

. l'indice de croissance de la consommation des facteurs $F_{1/F_0} = \dfrac{\sum_j p_j^0 \, f_j^1}{\sum_j p_j^0 \, f_j^0}$

...

[1] La rigueur des définitions - appuyée sur une formulation mathématique - ne doit pas être confondue avec les nécessaires conventions et approximations qui interviennent toujours dans la recherche de chiffres globaux.

On en déduit l'indice de croissance de la productivité globale.

$$1 + \Pi = \frac{P_{1/P_0}}{F_{1/F_0}}$$

Π est appelé *taux de productivité globale.*

. <u>Comparaison avec les productivités partielles</u>

L'indice de croissance de la consommation du facteur j est tout simplement le rapport des quantités consommées l'année 1 et l'année 0. L'indice de croissance de la productivité partielle du facteur j est donc :

$$1 + \Pi_j = \frac{P_{1/P_0}}{f_j^1/f_j^0}$$

Si on appelle $\varphi j = \frac{p_j^0 \; f_j^1}{F_1}$ le poids du facteur j dans le volume global des facteurs de l'année 1, on constate que :

$$1 + \Pi = \Sigma_j \quad \varphi_j \quad (1 + \Pi_j)$$

L'indice de croissance de la productivité globale est la somme des indices de croissance des productivités partielles des facteurs, pondérés par leurs poids respectifs dans le volume de facteurs de l'année 1.

On vérifie aisément que si les quantités des facteurs évoluent toutes de la même façon (on dit alors qu'on n'a pas de *substitution des facteurs*), $\Pi j = \Pi$ quelque soit le facteur j.

1.4. - <u>La dualité de la méthode</u> : <u>le surplus de productivité et sa répartition</u>

En augmentant sa productivité, l'entreprise a accru l'écart entre le volume des biens qu'elle restitue à la collectivité et le volume des facteurs de production qu'elle absorbe. Cet accroissement peut s'écrire :

$$S = \underbrace{(P_1 - F_1)} - \underbrace{(P_0 - F_0)}$$

Nous pouvons maintenant examiner comment se répartit le surplus de richesses ainsi créé. Pour cela, nous allons nous intéresser à la manière dont les prix varient.

Nous savons que les prix subissent une hausse générale due à l'inflation mesurée par l'indice du prix du Produit Intérieur Brut (P.I. B.). Pour chaque prix particulier nous pouvons évaluer sa dérive par rapport à la moyenne des autres : il suffit de raisonner en francs constants (en divisant les valeurs de l'année 1 par l'indice du prix du P.I.B. entre l'année 0 et l'année 1).

Quand le prix d'un produit diminue ou que celui d'un facteur augmente (en francs constants), c'est qu'une partie du surplus est utilisée à cet effet : on dit qu'on a un *emploi*. Dans le cas contraire, on a un *héritage* (ou une *ressource)*.

Le calcul en est simple. Soit p l'indice du prix du P.I.B.

Pour un produit : $q_i^1 \left(\frac{p_i^1}{p} - p_i\right)$ est un emploi s'il est négatif
est un héritage s'il est positif

Pour une charge : $f_j^1 \left(\frac{p_j^1}{p} - p_j\right)$ est un emploi s'il est positif
est un héritage s'il est négatif

p_i^1 et p_j^1 étant les prix des biens et facteurs l'année 1.

Cherchons la relation qui existe entre le surplus et les emplois et héritages.

Nous savons que :

$$S = (P_1 - F_1) - (P_0 - F_0)$$

ce que nous pouvons encore écrire :

$$S = \sum_i p_i^0 q_i^1 - \sum_j p_j^0 f_j^1 - \sum_i p_i^0 q_i^0 + \sum_j p_j^0 f_j^0$$

En introduisant les prix de l'année 1, nous obtenons :

$$S = - \sum_i q_i^1 \left(\frac{p_i^1}{p} - p_i\right) + \sum_j f_j^1 \left(\frac{p_j^1}{p} - p_j\right) \quad \left.\right\} \quad \textcircled{1}$$

$$+ \sum_i \left(q_i^1 \frac{p_i^1}{p} - q_i^0 p_i^0\right) - \sum_j \left(f_j^1 \frac{p_j^1}{p} - f_j^0 p_j^0\right) \quad \left.\right\} \quad \textcircled{2}$$

...

Nous reconnaissons dans le premier terme ① de l'équation la somme des emplois moins la somme des héritages ; regardons le deuxième terme. Il peut encore s'écrire :

$$\textcircled{2} = \frac{1}{p}\underbrace{\left(\sum_i q_i^1 p_i^1 - \sum_j f_j^1 p_j^1\right)}_{\text{bénéfice de l'année 1}} - \underbrace{\left(\sum_i q_i^0 p_i^0 - \sum_j f_j^0 p_j^0\right)}_{\text{bénéfice de l'année 0}}$$

On reconnait la variation en francs constants du bénéfice de l'entreprise.

On aboutit à l'égalité remarquable suivante :

$$\boxed{\text{surplus = emplois - héritages + variation du bénéfice}}$$

On constate qu'une entreprise qui augmente sa productivité n'accroît pas forcément son profit : par le jeu des emplois, elle peut en faire bénéficier ses clients, son personnel ou ses fournisseurs.
Productivité et profit ne sont en rien synonymes.

II - LES RESULTATS DE L'APPLICATION A EDF

1) Evolution de la productivité globale d'EDF de 1978 à 1979 - Répartition du surplus de productivité

...

Exprimés en francs 1979, les comptes d'exploitation des années 1978 et 1979 se présentent ainsi, après correction des charges de capital, et à hydraulicité normale :

. Tableau 1 : Comptes d'exploitation en valeurs (MF 1979)

	1978	1979
Ventes HT	7 120	7 994
Ventes MT	12 657	13 420
Ventes BT	23 732	25 382
TOTAL des produits	43 509	46 796
Combustibles fossiles	8 529	9 365
Combustibles nucléaires	768	1 227
Achats d'énergie	3 783	4 177
Personnel et institutions sociales	9 142	9 480
Dépenses diverses	5 919	6 317
Charges de capital (normatives)	27 735	28 562
TOTAL des charges	55 876	59 128
Résultat d'exploitation	- 452	- 9
Ecart entre les charges de capital effectives et normatives	11 915	12 323

...

A partir de ces données, on reconstitue des comptes d'exploitation en volumes : les valeurs des comptes 1979 sont exprimées, rubrique par rubrique, avec les prix (en F 1979) des comptes 1978.

. Tableau 2 : Comptes d'exploitation en volume (MF 1979)

	1978	1979
Ventes HT	7 120	7 923
Ventes MT	12 657	13 232
Ventes BT	23 732	25 380
TOTAL des produits	43 509	46 535
Combustibles fossiles	8 529	8 683
Combustibles nucléaires	768	992
Achats d'énergie	3 783	4 198
Personnel et institutions sociales	9 142	9 333
Dépenses diverses	5 919	6 301
Charges de capital (normatives)	27 735	28 562
TOTAL des charges	55 876	58 069

On en déduit le taux de P.G.F. :

$$1 + \pi = \frac{\dfrac{46\ 535}{43\ 509}}{\dfrac{58\ 069}{55\ 976}} = \frac{1{,}069}{1{,}039} \Longrightarrow \pi = 2{,}92\ \%$$

...

Le surplus de productivité est obtenu par comparaison des comptes en volumes des années 1978 et 1979. Les emplois et héritages sont calculés par soustraction des comptes de 1979 en valeurs et en volumes.

Parmi les bénéficiaires des emplois (ou les allocataires de ressources) figurent les prêteurs de capital. Il faut donc raisonner en charges de capital effectives et non normatives.

. Tableau 3 : Comptes d'exploitation de 1978 et 1979

(en volumes et valeurs, MF 1979)

	(1) 1978 (volumes-valeurs)	(2) 1979 (volumes)	(3) 1979 (valeurs)	(4) héritages	(5) emplois
Ventes HT	7 120	7 923	7 994	71	
Ventes MT	12 657	13 232	13 420	188	
Ventes BT	23 732	25 380	25 382	2	
TOTAL des produits	43 509	46 535	46 796		
Combustibles fossiles	8 529	8 683	9 365		682
Combustibles nucléaires	768	992	1 227		235
Achats d'énergie	3 783	4 198	4 177	21	
Personnel et institutions sociales	9 142	9 333	9 480		147
Dépenses diverses	5 919	6 301	6 317		16
Charges financières	15 820	5 273	4 129	1 144	
Amortissement		10 937	12 110		1 173
TOTAL des charges	43 961	45 717	46 805		
Résultat d'exploitation	- 452	+ 818	SURPLUS	1 270	
			Δ B		443
			TOTAL	2 696	2 696

...

Nous obtenons :

- le surplus de productivité S, par soustraction des produits et des charges de 1978 et 1979 en volumes :

$$S = \underbrace{(46\ 535 - 45\ 717)}_{\substack{\text{(produits-charges)} \\ \text{de l'année 1}}} - \underbrace{(43\ 509 - 43\ 961)}_{\substack{\text{(produits-charges)} \\ \text{de l'année 0}}} = 1\ 270 \text{ MF}$$

- la variation du bénéfice à francs constants, par soustraction des produits et des charges de 1978 et 1979 en valeurs :

$$B = \underbrace{(46\ 796 - 46\ 805)}_{\substack{\text{bénéfice de} \\ \text{l'année 1}}} - \underbrace{(43\ 509 - 43\ 961)}_{\substack{\text{bénéfice de} \\ \text{l'année 0}}} = 443 \text{ MF}$$

Vérifions que nous avons bien l'égalité :

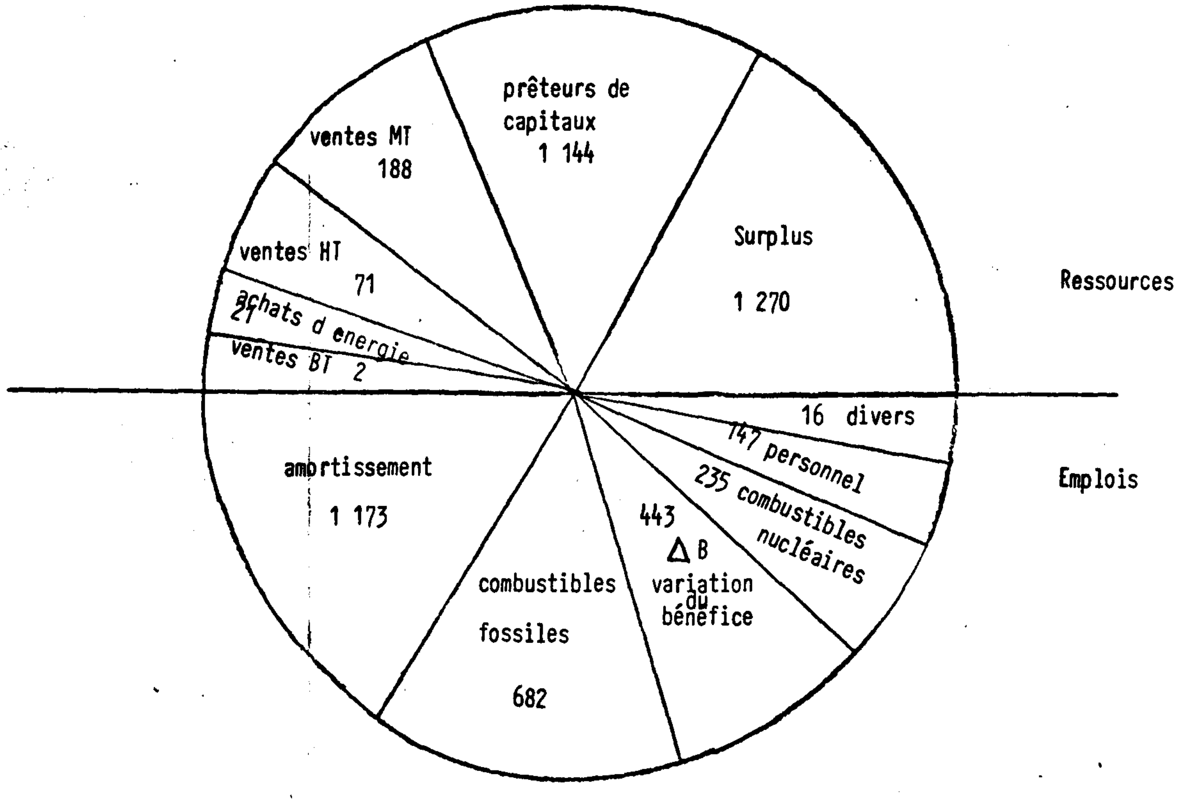
S = emplois - héritages + ΔB

Effectivement : 1 270 = 2 253 - 1 426 + 443

Cette égalité peut être représentée d'une manière qui permet de faire ressortir les rapports relatifs entre emplois et héritage (voir graphique) ce qui facilite l'interprétation.

. Répartition du surplus - interprétation

Somme du surplus et des héritages : 1 270 + 1 144 + 21 + 2 + 188 + 71 = 2 696 MF



Somme des emplois et de la variation du bénéfice : 682 + 235 + 147 + 16 + 1 173 + 443 = 2 696 MF

La plus grande partie des ressources vient des gains de productivité et des allègements des frais financiers versés aux prêteurs de capitaux (en l'occurence essentiellement l'Etat). Ces ressources sont venues compenser une hausse importante des prix des combustibles, et ont permis d'accroître l'autofinancement de l'entreprise (par le jeu de règles d'amortissement dégageant des provisions supérieures à la simple diminution de la valeur d'usage des actifs et par le rétablissement du résultat d'exploitation). La hausse des combustibles a été en partie répercutée sur les clients, qui n'ont donc pas bénéficié d'une partie du surplus dégagé (encore que, si le surplus avait été nul, la répercussion de cette hausse aurait sans doute été plus forte) au contraire du personnel.

2) Les progrès de la productivité globale des facteurs à EDF depuis 1960

Il est difficile d'interpréter la croissance de la productivité entre deux années consécutives, en partie parce qu'elle résulte de nombreux effets dont certains peuvent être conjoncturels, mais surtout parce que les décisions d'investissements sont prises en fonction du long terme, et donc que leurs conséquences sur la P.G.F. se font sentir sur plusieurs années. Pour toutes ces raisons, il est intéressant de retracer l'évolution des taux de P.G.F. sur une assez longue période, et d'analyser les tendances qui s'en dégagent.

Sur les deux graphiques ci-joints, nous avons reporté, outre l'évolution des taux de P.G.F., celle des volumes des produits et des facteurs (1).

Nous constatons que, d'un couple d'exercices à l'autre, le taux de P.G.F. varie notablement, donnant à la courbe une allure en dents de scie.

L'interprétation demande une certaine connaissance des données de la période. A la lumière de l'expérience, et en s'aidant du graphique des produits et charges, on peut cependant essayer de déceler certains phénomènes :

- au début des années 1960, l'accroissement annuel de la consommation a été important, permettant d'obtenir des taux de P.G.F. de l'ordre de 4,5 %, malgré une croissance sensible des charges ;
- de 1964 à 1968 environ, l'augmentation des charges a été sensiblement la même, mais celle des produits a été moins forte. Il s'ensuivit une dégradation de la P.G.F. qui s'explique essentiellement par le ralentissement de la consommation ;
- à partir de 1969 et jusqu'en 1975, alors même que la croissance de la consommation oscille autour de 8 % par an, le taux de croissance des charges ne cesse de baisser : on assiste en conséquence à un rétablissement de la P.G.F. ;
- ces dernières années, une nouvelle croissance des charges a ramené le taux de P.G.F. à un niveau plus bas.

...

[1] L'homogénéité des chiffres peut appeler quelques réserves en raison des modifications de calcul sur cette période, mais les ordres de grandeur et les évolutions sont corrects.

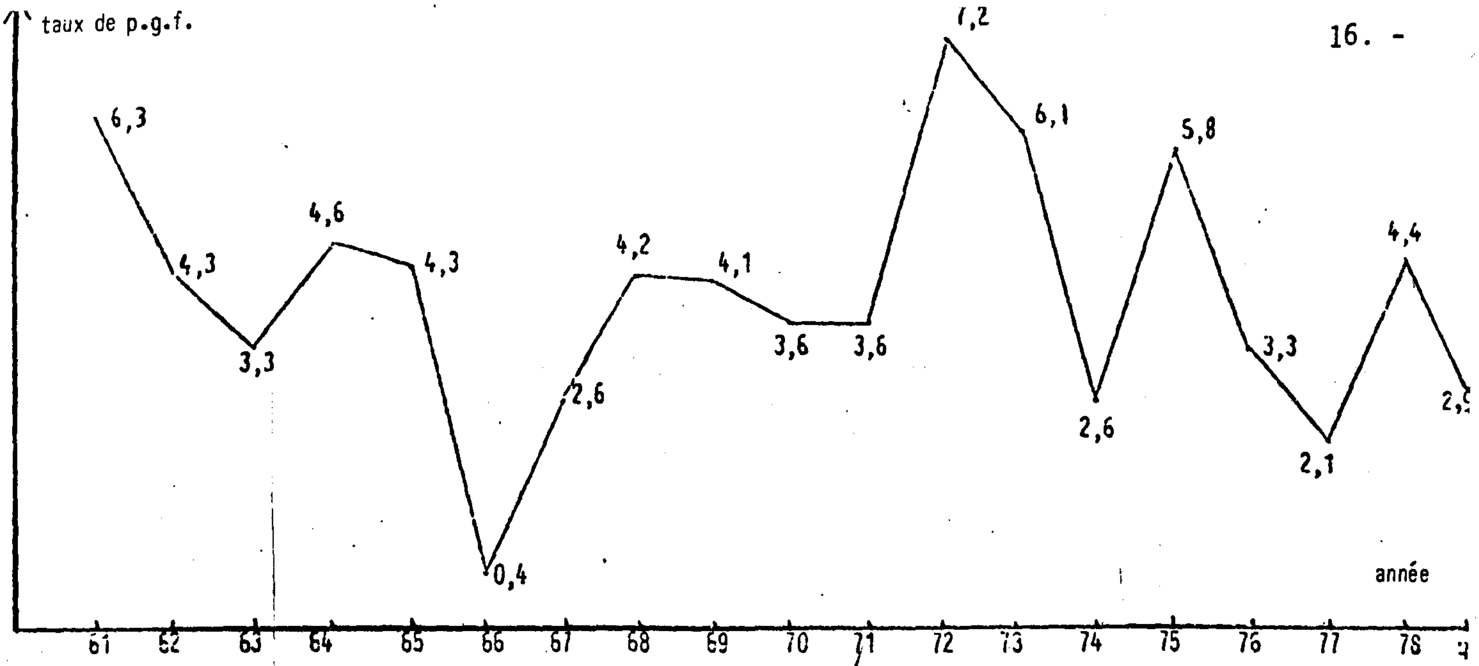


diagramme 1 : Evolution des taux de P.G.F.

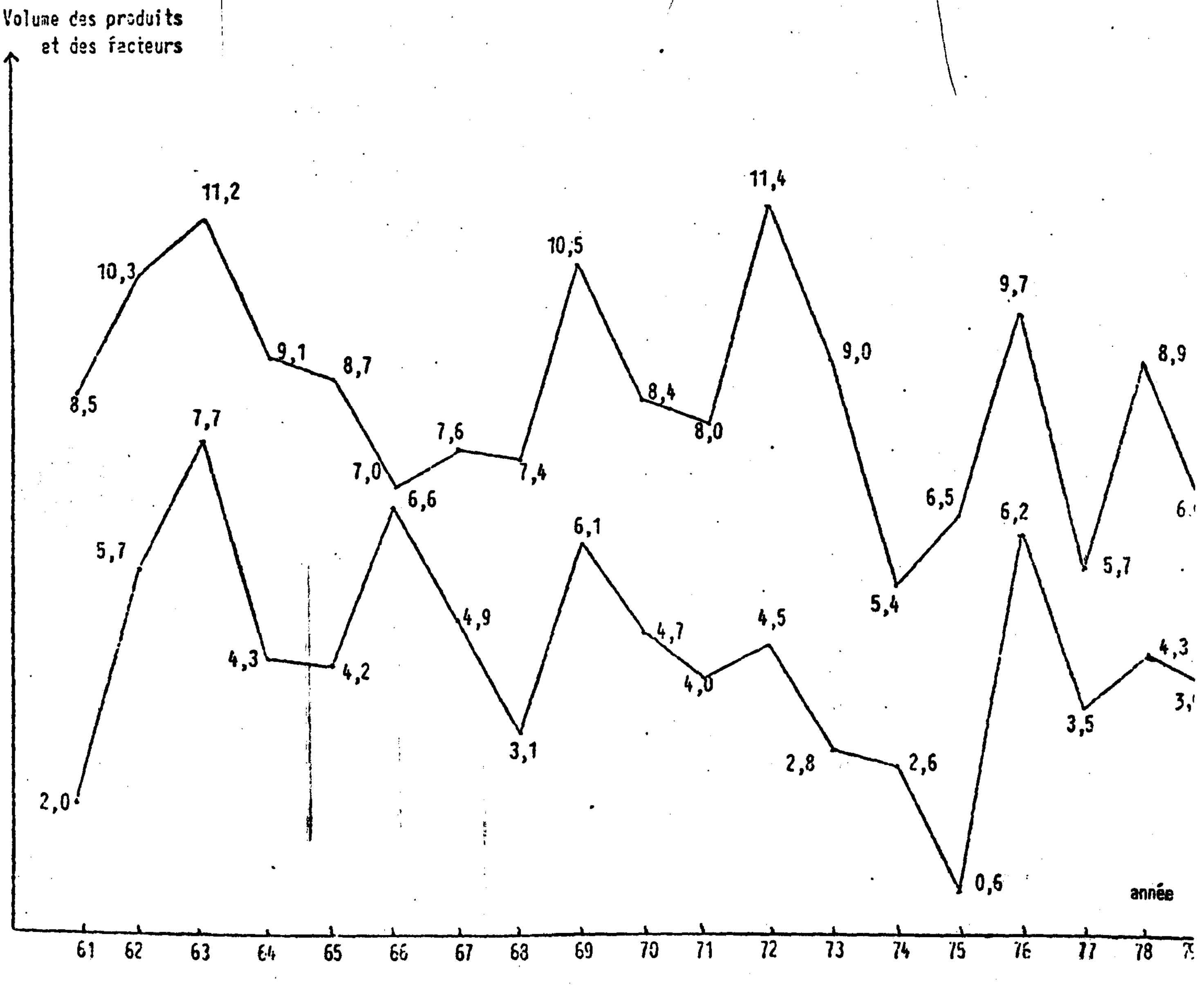


diagramme 2 : Evolution des volumes des produits et des facteurs

En moyenne, la P.G.F. a progressé de 4 % depuis 1960. Par delà les variations conjoncturelles, on peut donc affirmer qu'EDF est une entreprise dont la productivité est en nette croissance.

III - L'UTILISATION DE LA P.G.F. COMME OUTIL DE GESTION

1) Utilisation comme critère de jugement de la gestion d'EDF par les pouvoirs publics - L'expérience du contrat de programme

Depuis la nationalisation d'EDF, l'attitude des Pouvoirs Publics a progressivement évolué à son égard.

Dans les années d'après-guerre, les usines étaient détruites ; l'acier, le béton, le charbon manquaient. Il fallait reconstruire. Le vaste programme de relèvement national établi par le Premier Plan assignait une mission écrasante et urgente à toutes les entreprises publiques ; l'effort d'équipement nécessaire ne pouvait être assuré en ordre dispersé. C'est donc la logique d'un pouvoir centralisé qui a prévalu, comme cela se produit toujours en période de pénurie et de rationnement.

Les entreprises nationalisées des secteurs de base ont été utilisées directement et même "discrétionnairement" par les Pouvoirs Publics et, comme telles, elles ont été véritablement le moteur du redressement français. Car la centralisation, toujours dans le contexte de l'époque, favorisait la rationalisation des moyens et des méthodes industriels et la reconstitution rapide des infrastructures.

Les problèmes techniques passaient après la politique conjoncturelle dont le secteur public, du fait de son poids massif dans l'économie, allait devenir très vite un instrument privilégié. Quoi de plus tentant pour un Gouvernement ? L'action sur les prix et les salaires, les contraintes tarifaires imposées en vertu d'objectifs sociaux ou régionaux avaient évidemment, concernant un secteur aussi vaste, de puissants effets à court terme sur l'économie.

Mais la gestion d'une entreprise, qu'elle appartienne au secteur public ou au secteur privé, et surtout lorsqu'elle est très capitalistique, ne peut s'accommoder longtemps de décisions qui ont un autre objet que la vocation même de l'entreprise. Si bien des entreprises nationales ont rencontré, au fil des années, des difficultés de gestion certaines, c'est dans la mesure où elles ont servi à autre chose qu'à leur destination première.

Face à cette situation, a été créé en 1967, à la demande du Premier Ministre, un groupe de travail interministériel dont le rapport devait être largement utilisé dans l'élaboration des "contrats de programme".

Ce rapport mettait l'accent sur les problèmes de financement du secteur public, notamment dans le domaine de l'énergie et du transport où les entreprises, largement capitalistiques, doivent faire face à des investissements très lourds. Il dégageait l'idée que, dans le nouveau contexte industriel, il était nécessaire que l'Etat renonce à utiliser ces entreprises à des fins de politique économique générale, cette pratique se faisant au détriment de l'autonomie des entreprises concernées et sans aucun avantage réel pour la puissance

...

publique. Il préconisait un retour à une conduite plus saine de ces entreprises qui devait leur permettre de dégager les moyens nécessaires à leur développement et au Service Public sans faire appel de façon massive aux deniers de l'Etat ou au marché financier. Il faisait valoir que, dans un contexte d'équilibre économique retrouvé, la tâche essentielle des entreprises nationales n'était plus d'atteindre, coûte que coûte, des objectifs physiques, mais d'améliorer leur productivité. Or, indiquait le rapport, l'efficience n'est pas compatible avec une gestion centralisée du secteur nationalisé : "Seule l'entreprise a de son marché et de ses propres moyens une connaissance suffisante pour pouvoir élaborer et appliquer une politique efficace de productivité et de compétitivité. Force est ainsi d'accroître son autonomie, quitte à sanctionner sa gestion au vu des résultats obtenus". L'idée du contrat de programme était née. Mais, ajoutait le rapport, "la conduite efficiente de l'entreprise publique plus autonome et mieux orientée après son marché implique des modalités internes de gestion qui ne sont pas toutes encore réunies dans de nombreuses entreprises publiques".

Qu'en était-il à Electricité de France ? Rappelons-le, le texte de la loi de nationalisation était très discret sur tout ce qui concernait la gestion de la nouvelle entreprise.

Mais, soucieuse de l'intérêt de la collectivité, l'entreprise a peu à peu fixé les critères de son action autour de quatre principes essentiels que nous rappelons : l'égalité de traitement des usagers devant le Service Public ; la vente au coût marginal (en développement) ; le choix des investissements par référence à un coût de défaillance pour la collectivité et par référence au taux d'actualisation du Plan ; enfin, le critère synthétique de gestion conduisant à maximiser les progrès de productivité globale des facteurs.

Dans ce cadre, il était possible de rationaliser les choix techniques et économiques, problème majeur pour une industrie qui investit chaque année l'équivalent d'environ 40 à 50 % de son chiffre d'affaires. De fait Electricité de France, qui n'a jamais été déficitaire en tendance et qui a toujours fait preuve d'une très forte productivité, constituait un terrain favorable et réunissait les meilleures conditions pour négocier un contrat de programme avec l'Etat.

Le contrat de programme à EDF

Si Electricité de France avait déjà acquis ces résultats, pourquoi un contrat de programme ? Les performances réalisées par cet établissement prouvaient qu'une certaine harmonie existait dans les relations entre l'Etat et l'entreprise. Cela est tout à fait vrai et cette entente préalable avec l'Administration a beaucoup facilité les choses lors de l'élaboration du contrat de programme signé en décembre 1970. Le dispositif contractuel comporte deux documents indissociables : une "lettre de mission" adressée au Président du Conseil d'Administration d'EDF sous la signature du Ministre du Développement Industriel et Scientifique et du Ministre de l'Economie et des Finances (ce document unilatéral précise la politique du gouvernement à l'égard de l'établissement), et le contrat de programme signé par les deux

...

parties et qui définit leurs engagements réciproques pour la période couverte par ce contrat (1971-1975).

La lettre de mission constitue en quelque sorte la charte des relations nouvelles que l'Etat entend établir avec l'entreprise.

Le gouvernement y consacre la vocation industrielle et commerciale d'EDF : il reconnaît la nécessité, face aux développements d'une économie désormais ouverte à la concurrence internationale, "de renforcer les moyens" dont dispose l'établissement pour assurer sa gestion conformément à cette vocation ; il retient explicitement parmi les efforts que doit poursuivre l'entreprise pour accomplir sa mission "le développement d'une action commerciale efficace" ; il affirme qu'Electricité de France, jugée désormais sur la réalisation des objectifs qui lui sont fixés par le contrat, "sera responsable du choix des moyens pour les atteindre" ; il accepte qu'après entente sur un programme d'investissement étalé sur cinq ans, le montant global en soit retenu à titre d'orientation, l'entreprise conservant la possibilité d'investir plus si ses ressources propres le lui permettent ; il recommande à EDF de chercher à harmoniser sa politique d'équipement et d'achat à la politique industrielle des Pouvoirs publics ; tout en lui reconnaissant la responsabilité finale de la décision, il accepte, sauf à respecter certaines recommandations, que l'établissement prenne les participations ou entreprenne les activités nouvelles utiles à l'exercice de son activité principale et à la promotion commerciale de son marché ; il précise qu'entendant laisser désormais à l'entreprise la responsabilité de ses décisions, le contrôle de la gestion d'Electricité de France s'effectuera a postériori.

En clair, ce dispositif concentre sur l'entreprise des responsabilité antérieurement très diluées au sein de l'Administration. Quels étaient, dans le contrat proprement dit, les objectifs fixés à Electricité de France et les contreparties consenties par l'Etat ? Il était demandé à l'entreprise d'atteindre, sur une période de 5 ans, une productivité globale plus élevée encore que par le passé (le taux moyen est fixé à 4,85 %, à comparer avec celui de l'industrie française qui est d'environ 3 %), un autofinancement accru, un haut niveau de rentabilité financière du capital investi. Du côté des Pouvoirs publics, on trouvait en contrepartie l'octroi d'une plus grande autonomie en matière de tarifs (réformes de structures, mise à niveau...) et de prises de participation, un engagement pour deux ans sur le concours de l'Etat et les possibilités d'accès d'EDF au marché financier en vue d'assurer le financement des investissements, et enfin une déclaration de principe sur l'allégement des mesures de contrôle (contrôle a priori, autorisations préalables).

Le contrat s'est soldé par un succès certain, puisque l'objectif de productivité a été largement atteint (5,1 % de croissance annuelle moyenne de la PGF sur la période) et qu'EDF a apprécié l'autonomie dont elle a bénéficié.

A partir de 1975, la nécessité d'adapter le parc de production aux nouvelles conditions économiques qui ont suivi la crise pétrolière - et les problèmes de financement induits - ont prévalu sur les préoccupations d'amélioration de la productivité. Mais, ces dernières reviennent aujourd'hui à l'ordre du jour, les problèmes de production étant en voie d'être réglés

...

avec la reconversion des centrales du fuel en charbon et la mise en service des nouvelles tranches nucléaires.

2) Utilisation comme outil de gestion prévisionnelle - les problèmes d'interprétation

Chaque année, EDF établit une chronique de comptes d'exploitation prévisionnels sur le moyen terme (de l'ordre de 10 ans), qui lui permet notamment d'avoir une idée de ses besoins de financement futurs, et une base de discussion avec les Pouvoirs publics sur l'évolution des niveaux tarifaires. Pour chaque poste, des hypothèses de productivité (partielles) sont faites, et, sur la base des comptes ainsi obtenus, un taux de productivité globale est calculé, qui permet d'avoir une idée synthétique des progrès de productivité escomptés.

L'interprétation de ce taux n'est cependant pas toujours aisée.

Certes, nous l'avons vu, le taux de PGF est une pondération de taux partiels représentatifs des différents facteurs. De la même façon, l'indice de volume des produits est une pondération des taux de croissance des différents biens et services produits par l'entreprise. Lorsque les poids relatifs des facteurs et des produits varient peu, le taux de PGF reflète directement l'accroissement des productivités partielles. Mais il en va autrement dans les périodes où, soit les prix des facteurs varient notablement les uns par rapport aux autres, soit les évolutions technologiques ou de structure de la consommation entraînent une modification des volumes relatifs des produits et facteurs. Par définition, un taux de PGF "englobe" tous les phénomènes qui peuvent agir sur lui, et il est parfois difficile d'isoler leur action spécifique (leur "effet").

- Effet qualité de service

Nous l'avons déjà évoqué plus haut au sujet de la non-valorisation des biens non-marchands. Rappelons qu'en fonction de l'équipement de production, de transport et de distribution, pour un appel de consommation donné, les temps de coupure et les chutes de tension sont plus ou moins importants. Il faudrait donc affecter aux kWh non desservis (et aux baisses de tension) un prix normatif qui représente la gêne qu'ils entraînent pour la collectivité.

Nous avons évoqué la difficulté de sa prise en compte dans le calcul ex-post. Dans le calcul ex-ante, la "correction" envisagée trouve son sens : elle permet d'éliminer cet effet, de façon à recalculer un taux de P.G.F. "pur", c'est à dire comparable à des situations présentant la même qualité de service. Elle s'effectue, lorsqu'il existe un déficit de puissance garantie par rapport à l'ajustement (1), en ajoutant aux charges de capital le coût nécessaire pour installer et exploiter (hors combustible) un équipement de pointe garantissant cette même puissance.

...

(1) L'ajustement consiste à construire des équipements (dits "de pointe") de faible coût d'installation jusqu'à ce que leur coût d'investissement et d'exploitation soit égal au coût de la défaillance évitée.

Une remarque similaire pourrait être faite à propos des dépenses engagées par l'Etablissement au titre de l'environnement. Dans la mesure où elles accroissent le coût des ouvrages sans qu'on ne chiffre explicitement dans les produits l'intérêt correspondant pour la collectivité, elles ont une influence dans le sens de la baisse du taux de productivité calculé, qu'il serait souhaitable mais difficile de corriger.

2) Effet de substitution des facteurs

La rationalité économique impose à une entreprise de combiner ses facteurs de production de façon à minimiser son coût total de production. A un rapport de prix donné des facteurs correspond une combinaison donnée. Si ce rapport change, l'entreprise a intérêt à changer de combinaison, de façon à utiliser une quantité moindre du facteur dont le prix a augmenté.

Cette adaptation à l'évolution des prix se fait de manière plus ou moins rapide suivant qu'elle nécessite des investissements plus ou moins importants. C'est ainsi que, suite au quadruplement du prix du pétrole en 1973-1974, la réaction de l'Etablissement s'est faite en trois étapes :

- d'abord par une utilisation plus intensive des centrales au charbon, dont les effets sont immédiats ;

- ensuite par une reconversion au charbon d'un certain nombre de centrales fonctionnant au fuel ;

- enfin par une substitution progressive du nucléaire aux combustibles classiques, qui commence à porter ses fruits.

Chaque substitution diminue la quantité (pondérée avec le nouveau système de prix) des facteurs, alors même que les techniques utilisées peuvent être les mêmes (ou même, rapportées à chaque facteur, plus chères : le coût total d'une centrale au fuel reconvertie au charbon est supérieur à celui d'une centrale construite pour brûler du charbon).

La PGF augmente. On dit que l'on a un effet de *substitution de facteurs*. Cet effet s'annule lorsqu'on s'est réadapté à la nouvelle structure des prix.

On peut donner une mesure de l'effet de la substitution du nucléaire aux combustibles classiques, au moins de manière approximative : les comptes de combustible et de capital sont récalculés en admettant que la même demande est satisfaite (les recettes sont donc inchangées) sans substitution, et en faisant les investissements appropriés dans les moyens de production classiques pour obtenir la même puissance garantie (et donc la même qualité de service).

La différence entre le taux de PGF obtenu précédemment et le nouveau taux ainsi calculé donne une mesure de l'effet nucléaire.

...

3) Effet d'expansion

L'une des caractéristiques les plus marquantes de la distribution de l'électricité est que le volume de capital et de personnel nécessaire pour desservir la clientèle augmente moins vite que la consommation : à conditions techniques identiques, plus la densité de puissance à desservir est grande, plus bas est le coût de distribution. On dit que les rendements sont croissants.

Au niveau de la production, les rendements sont croissants mais beaucoup plus faiblement en production thermique, grâce aux effets de taille et au foisonnement (plus il y a d'unités de production, moindre est la conséquence de la panne d'une unité). Ils deviennent décroissants en production hydraulique, par suite de la saturation des sites exploitables (on dit qu'on a un "facteur externe" défavorable).

Néanmoins, compte tenu de la forte proportion des basse et moyenne tensions, à une croissance de la consommation correspond une augmentation *automatique* de la productivité : c'est *l'effet d'expansion.*

A cet effet se superpose un effet de *structure* des produits (analogue à l'effet de substitution des facteurs). Il n'est pas indifférent pour le taux de PGF que la BT augmente plus vite que la HT : l'effet d'expansion joue plus sur la première que sur la seconde.

4) Effet de poids des facteurs

Nous avons vu que le taux de PGF pouvait s'écrire comme une somme pondérée par les poids des facteurs des productivités partielles. Si ces dernières sont très différentes suivant les facteurs, les poids vont avoir une influence très importante. Or les différences existent : en l'état actuel de la technologie, il n'est pratiquement plus possible d'améliorer le rendement des centrales thermiques ; la productivité partielle du facteur "combustibles" est donc constante. Or le poids de ce facteur est directement fonction de son prix, qui est en constante variation : le taux de PGF va donc en être directement affecté. C'est l'effet de *poids des facteurs.*

5) Mesure des progrès de gestion courante

Bien qu'il soit un peu périlleux de vouloir isoler des effets qui sont relativement dépendants, ils n'en reste pas moins qu'une fois ramené à une qualité de service égale, et une fois corrigé l'effet d'expansion et de poids des facteurs, il est intéressant d'isoler l'effet substitution (qui dépend, notamment en ce qui concerne la substitution du charbon et du nucléaire au fuel, d'une politique centrale de l'Etablissement) du taux pour faire apparaître l'effet de phénomènes diffus, dépendants, et induits par des décisions décentralisées, tels que l'évolution générale de la technologie, la meilleure organisation du travail, l'apprentissage des techniques employés,...

Il ne faut pas cependant perdre de vue que la validité du taux résiduel ainsi obtenu dépend fortement de celle de la mesure des différents effets qu'on a isolés. On obtient plus un ordre de grandeur qu'une valeur véritablement fiable.

CONCLUSION

Le taux de productivité globale des facteurs constitue un instrument d'analyse significatif de la qualité de la gestion d'un Service Public : il est homogène aux critères de décision conformes à l'intérêt général.

Utilisé depuis déjà longtemps à EDF, il permet de suivre la performance de l'Etablissement au cours des ans. C'est un critère de gestion synthétique irremplaçable pour une politique contractuelle qui allie l'autonomie de gestion concédée à une entreprise de Service Public à son nécessaire contrôle par les pouvoirs publics. Utilisé avec succès lors d'un premier Contrat de programme, il est très probable qu'il serait de nouveau inscrit à un nouveau Contrat, si ce dernier venait à être signé.

On lui reproche parfois d'être d'application malaisée et d'interprétation difficile. Pourtant, nous l'avons vu, une bonne connaissance des phénomènes économiques qui le sous-tendent permettent de l'analyser, et des procédures de calcul relativement simples donnent la possibilité de mettre en évidence les effets les plus significatifs. Il ne donne donc pas seulement une mesure <u>globale</u> de la performance - en quoi il est irremplaçable - mais grâce à une analyse plus fine on peut mesurer l'impact spécifique des décisions importantes de l'entreprise.

- 0 -

- 0 -   - 0 -

NET INCOME AND PRODUCTIVITY ANALYSIS (NIPA) AS A PLANNING MODEL

ALI CHAUDRY

MALCOLM BURNSIDE

American Telephone and Telegraph Company

Abstract

Most planning models and the conventional financial analyses of a firm's performance focus on sales, costs and profits, which include the effects of price changes as well as changes in physical volumes over time. As these analyses do not consider productivity changes explicitly, they could mask a decline in efficiency of the enterprise or fail to provide correct indications of any productivity improvements that might be taking place. A Net Income and Productivity Analysis (NIPA) model has been developed to provide a link between conventional financial measures and productivity. NIPA quantifies the dollar contribution of the various determinants of growth in net income from one year to the next, and separates these into price and quantity components. The basic NIPA model can be used for historical analysis of performance as well as for corporate planning and budgeting for future years. This paper develops several extensions, which permit evaluations of prespecified financial targets with explicit measures for productivity, capital growth, inflationary cost increases and changes in depreciation and taxes.

Key Words: Net Income, Total Factor Productivity, Financial Analysis, Budgeting, Corporate Planning.

I. INTRODUCTION

The purpose of this paper is to describe a productivity-based planning model which is in current use at a major U.S. corporation and which explicitly recognizes the role total factor productivity plays, along with other factors in determining the bottom line net income. Using hypothetical data for XYZ Corporation, we illustrate how an existing planning and budgeting process can be enhanced by introducing the contribution of productivity and other factors in terms of dollars and cents to which decision-makers can relate easily.

Conventional financial analyses of a firm's performance tend to focus on sales, costs and profits in nominal terms which include the effects of price changes as well as changes in physical volume of outputs and inputs. Similarly, many planning models deal with nominal variables, relying almost exclusively on financial ratios as the basis for planning decisions. For example, Steffy et al (1974) suggest using a multiple regression equation to explain and project net income as a function of ratios such as sales-to-inventory, current assets-to-current liabilities, sales-to-capital funds, sales-to-net worth, etc., and in terms of variables such as the average collection period and the "Acid Test Ratio," defined as the ratio of cash to current liabilities. Models that do consider the physical side of the picture, often fail to take productivity into account explicitly. Thus they might mask a decline in efficiency of the enterprise and fail to provide accurate signals for management to take corrective action. By the same token, these models could fail to detect opportunities for improvement in efficiency that management should know about.

To rectify these deficiencies, a Net Income and Productivity Analysis (NIPA) Model (Chaudry, Burnside and Eldor (1980)) has been developed to explain the growth in net income in terms of total factor productivity (TFP), growth of capital, price changes, inflation in input costs, depreciation, taxes and other financial factors*. We refer to these as NIPA factors. The NIPA model assumes total product exhaustion by inputs and other specific factors in each period and attributes dollar values to all inputs and factors. By regrouping them according to the way they affect net income, the net sum of these factors accounts for the total change in net income.

These factors can be grouped into the following categories:

Income-Augmenting Factors - those directly contributing to growth in net income: productivity, or the improvement in efficiency of the firm; growth in the physical capital stock; changes in product prices; and 'other income' (not directly associated with the physical operations of the firm).

* Werner (1979) has also developed a similar productivity-based model which is designed to calculate a theoretical budget, subject to appropriate constraints facing the corporation. NIPA, by contrast, works with the proposed budget and recasts it in terms of productivity and other factors discussed in the text.

Income-Absorbing Factors - those inversely related to growth in net income: changes in prices of materials and services purchased from other firms; changes in labor input costs due to changes in wages and benefits; changes in non-income (indirect) taxes due to change in the tax rates; changes in depreciation expenses; and changes in income taxes and financial factors. (The financial factors included in this category are interest charges, uncollectibles, miscellaneous deductions from income and extraordinary and delayed items - net. In the detailed model output, these factors are analyzed individually. See Table 1.)

The NIPA results can also be presented in the form of an "arrow chart" (See Figure 1). The length of each arrow shows the magnitude of the impact of each factor upon the change in net income for the year under study and the point of the arrow indicates the direction of the impact. The income augmenting factors are shown first in a cumulative fashion, followed by a cumulative netting out of the income absorbing factors, with the difference exactly matching the change in net income for the year. Note, however, that one of the income absorbing factors in this figure - Non-Income Taxes - acted to increase net income, since in this year these taxes were actually lower relative to the real capital stock and revenue (on which they are based) than they were in the previous year.

For brevity, a number of financial factors have been combined with Income Taxes and shown on the arrow chart as "Tax and Financial" factors. Of course, if desired, each of the underlying variables could be shown separately in this figure, including components of the productivity calculation. However, the detail might clutter the chart; if a picture is supposed to be worth a thousand words, it had better be crisp and clear.

This model can be used by a corporation for corporate planning, budgeting and for strategic targetting to develop and analyze its budget or its corporate plans in such a way that (1) all relationships between productivity and costs, and those between prices and volume of business are treated consistently and (2) all financial factors are fully accounted for. In other words, nothing is allowed to fall through the cracks, as is common with many financial models that rely on ad hoc ratios of selected variables. For example, these ratios are usually calculated in terms of current prices and thus reflect the effects of both price changes and volume changes. NIPA, on the other hand, decomposes each key variable into its price and quantity components and ensures that the total change in the variable is accounted for by the sum of the changes in the respective components.

We describe the theoretical framework of the model in Section II followed by a discussion in Section III of how the basic results might be analyzed in terms of the dollar impact of changes in prices and quantities separately, on the change in net income.

This feature of NIPA is the key to its usefulness as a planning model. The user can vary any of the assumptions explicitly in terms of projected productivity gain, price changes, inflation in various costs and other factors such as tax rates or depreciation rates. For illustrative purposes, Section IV describes three alternative scenarios based on different productivity projections. Some concluding remarks are presented in Section V.

II. THE MODEL

By the usual broad definition, net income is simply the difference between total sales revenue and total expenses or costs. Defining revenue as all financial inflows and expenses as all financial outflows, including taxes etc., we have

$$NI = R - C^* \tag{1}$$

where

NI = Net Income
R = Total Sales Revenue
C^* = Total Operating Costs (excluding return to equity capital).

Thus, change in net income can be expressed as:

$$dNI = dR - dC^* \tag{2}$$

where d indicates a discrete change in the respective variable from one year to the next, measured in dollars as shown on the income statement. Time subscripts have been omitted here for simplicity of notation and will be used subsequently as needed.

Since the changes in revenue and costs, and therefore in net income, reflect the combined effect of price and quantity changes, we need to further decompose the total change in each variable into its price and quantity components. Only then can we measure productive efficiency in terms of the real output and real inputs and account for the price effects separately.

It should be noted that the productivity calculation in NIPA is different from the fixed base-year methodology which is used in standard TFP measurement. Since we are attempting to account for the growth in net income from one year to the next, we are dealing with the quantities (or volumes of output and inputs, respectively), and their corresponding prices for two consecutive years. Thus, the measures of output and inputs for the current year (t) must be constructed in terms of prices of the previous year (t-1). Similarly, any effects attributable to changes in output prices or input prices must be measured with reference to the previous year. This means that all computations of this type to be made within the NIPA framework employ a changing base year as contrasted with the fixed-base-year indexes of traditional productivity studies.

If we think of net income as a return to equity holders, and thus a claim against total revenues, total costs might be defined as

$$C = C^{*} + NI. \tag{3}$$

We may then write the model requirement of total revenue exhaustion as an equality between total revenue and total cost, or

$$R = C. \tag{4}$$

It then follows that

$$dR = dC \tag{5}$$

where

$$dR = P(Q).dQ + dP(Q).Q \tag{6}$$

with P(Q) representing the base year price of base year output Q; and

$$dC = P(X).dX + dP(X).X \tag{7}$$

where P(X) is a vector of prices corresponding to the input vector X.

Thus, we have

$$P(Q).dQ + dP(Q).Q = P(X).dX + dP(X).X \tag{8}$$

where

$$P(Q).dQ \text{ and } P(X)dX \tag{9}$$

represent the real physical effects and

$$dP(Q).Q \text{ and } dP(X).X \tag{10}$$

represent the effect of changes in output and input prices respectively. By definition, total productivity (TP*) is

$$TP = \text{Change in Real Output} - \text{Change in Real Inputs} \quad (11)$$

or

$$TP= P(Q).dQ - P(X).dX. \quad (12)$$

The accounting identity (8) above can be written as

$$P(Q).dQ + dP(Q).Q - P(X)dX - dP(X).X = 0. \quad (13)$$

Separating the real and price components, we can write it as

$$P(Q).dQ - P(X).dX + dP(Q).Q - dP(X).X = 0 \quad (14)$$

or by using the definition of TP in (11) above,

$$TP + dP(Q).Q - dP(X).X = 0. \quad (15)$$

* TP as used here refers to the total productivity of all inputs, namely capital, labor and materials. The more commonly known concept of total factor productivity (TFP) is sometimes used interchangeably with TP. However, some authors prefer to limit the use of TFP to the combined productivity of capital and labor only.

Alternatively,

$$TP = dP(X).X - dP(Q).Q \tag{16}$$

which is a definition of TP in terms of output and input prices. Substituting individual input prices explicitly, we can write the total change due to changes in these prices as

$$dP(X).X = dP(K).K + dP(L).L + dP(M).M \quad . \tag{17}$$

where K is total capital, L is labor and M is total materials, rents and services.

The identity (15) can now be written as

$$TP + dP(Q).Q = dP(K).K + dP(L).L + dP(M).M. \tag{18}$$

Equation (18) states that the total productivity gain and the value of the output price changes are absorbed by the three inputs in the form of remuneration to the respective factors of production.

The foregoing exposition has been simplified by explicitly including only the quantities and prices of the three major input factors. We also need to take into account indirect taxes and a number of financial factors for completeness. In defining output for calculating TP, deflated indirect non-income taxes (NIT) are generally subtracted from deflated revenues. These include (a) Property taxes, (b) Capital Stock taxes, (c) Gross

Receipts taxes, and (d) other Non-income taxes. The first two categories are related to the real investment in plant and equipment, while the last two are related to sales revenue and these relationships are used to make the calculations.

The total change in these taxes (dNIT) consists of real change (dNITR) and the "price change" effect which, in this case, means the change resulting from a change in the tax rate (dNITP), i.e.,

$$dNIT = dNITR + dNITP. \qquad (19)$$

The real effect, dNITR has been implicitly accounted for in the definition of output and of TP*, as given above, but we also need to account for the "price effect." This is done by expanding the (dP(X).X) vector to add dNITP to the right hand side (RHS) of equation (18).

Since the capital input change is a deduction in the TP calculation, but is included (in part) in the net income, we need to reflect this in our model.

* In terms of tax-adjusted output and real inputs, total productivity is defined as:

$$TP = \underset{\substack{\text{Total}\\\text{Deflated}\\\text{Revenue}}}{P(Q).\ dQ} - \underset{\substack{\text{Deflated}\\\text{Non-Income}\\\text{Taxes}}}{dNITR} - \underset{\substack{\text{Real}\\\text{Capital}\\\text{Input}}}{P(K).dK} - \underset{\substack{\text{Labor}\\\text{Input}}}{P(L).dL} - \underset{\substack{\text{Materials}\\\text{Input}}}{P(M).dM.}$$

Adding P(K).dK to both sides of the equation, we get

$$\begin{aligned} TP + dP(Q).Q + P(K).\ dK = {} & (dP(K).K + P(K).dK) \\ & + dP(L).L + dP(M).M + dNITP \qquad (20) \end{aligned}$$

where P(K).dK is the growth of physical capital input and dP(K).K + P(K).dK = d(P(K).K) is the current undeflated value of the change in capital input. For the present expository purpose the latter may be interpreted as comprising depreciation (DEP), interest charges on debt (INT), income taxes (FIT + SLIT), and the return to equity investors (i.e., the net income (NI) - including other income (OI)) and other miscellaneous financial factors such as uncollectibles (UNC), miscellaneous deductions from income (MDI) and extraordinary and delayed charges and credits-net (E&D).

Substituting these factors for d(P(K).K), we obtain the fundamental equation underlying NIPA as

$$\begin{aligned} TP + dP(Q).Q + P(K)\ dK + dOI = {} & dP(L).L + dP(M).M + dNITP \\ & + dDEP + dFIT + dSLIT \\ & + dINT + dUNC + dMDI \\ & - dE\&D + dNI \qquad (21) \end{aligned}$$

This equation is an alternative definition of the change in net income which we present on the following page, along with the traditional definition contained in the income statement. This equation is also the basis for all simulation results generated by NIPA which we discuss in subsequent sections of this paper. The complete set of relationships described in this section is shown schematically in Figure 2, with all variables as defined in the text. (For a compact list of these definitions, see the Appendix.)

The whole model can be thought of as consisting of four interrelated modules, namely, Productivity Module, Capital Growth Module, Price Effects Module and finally Tax and Financial Module. We will return to the relationships among these modules when we discuss the alternative planning scenarios in Section IV.

Relationship Between NIPA and the Income Statement

		NIPA		INCOME STATEMENT
		Change In Net Income =		Change in Net Income =
		Change in Deflated Revenue		Change in Revenues[1]
	-	Change in Def. Non-Income Taxes	-	Change in Non-Income Taxes[1]
TP	-	Change in Labor Input	-	Change in Labor Costs[1]
	-	Change in Capital Input	-	0.0[2]
	-	Change in Materials Input	-	Change in Materials Costs[1]
	+	Output Price Changes		(See Footnote 1)
	+	Capital Expansion	+	0.0[2]
	-	Inflation in Materials		(See Footnote 1)
	-	Inflation in Labor		(See Footnote 1)
	-	Inflation in Non-Income Taxes		(See Footnote 1)
	+	Change in Other Income	+	Change in Other Income
	-	Change in Depreciation	-	Change in Depreciation
	-	Change in Federal Income Tax	-	Change in Federal Income Tax
	-	Change in State & Local Income Tax	-	Change in State & Local Income Tax
	-	Change in Interest	-	Change in Interest
	-	Change in Uncollectibles	-	Change in Uncollectibles
	-	Change in Misc. Deductions	-	Change in Misc. Deductions
	+	Change in Extra. & Del. Items-Net	+	Change in Extra. & Del. Items-Net

Notes:

1. These items are in nominal terms and thus include price changes.
2. In the income statement, there is no deduction for capitalized investment expenditures. Thus the return to capital is a part of net income.
3. TP = Total Productivity. See text for definition.

Differences between NIPA and Conventional TFP Studies

NIPA differs from the conventional TFP or TP studies in several respects. Most conventional productivity studies use a fixed base year for purposes of developing the necessary quantity and price indexes. While NIPA utilizes some fixed-base-year deflators in certain preliminary calculations, the essential calculations are made by continuously shifting the base year for each index. Specifically, for any pair of years being analyzed, NIPA treats the first as the base year for calculating quantities (volumes) or quantity indexes for the second year. Thus the quantity indexes are meaningful for only that pair of years and cannot be used to compute growth rates over longer periods.

This concept of the shifting base year also applies to the unit price of capital which changes every year in NIPA, while it is fixed at the base-year price in conventional studies.

In many respects, the NIPA procedures are similar to the Divisia index methodology used in many recent TFP studies*. The current version of the NIPA model, however, does not incorporate shifts in all weights as

* For example, see M. Denny, M. Fuss and L. Waverman, "The Measurement and Interpretation of Total Factor Productivity in Regulated Industries with an Application to Canadian Telecommunications," a paper presented at the NSF conference on Productivity Measurement in Regulated Industries, Madison, Wisconsin, 1979. This paper also contains an excellent list of other references to the Divisia literature.

required by Divisia indexes. For instance, in the current version of the model total weighted hours are calculated with fixed relative weights as of a particular base year (as in traditional studies). It is certainly possible to adopt shifting weights at this level of the calculations, although it is not clear that such a refinement will have a substantial effect on the results, especially if the firm has fairly stable labor weights and a stable work force mix.

III. ANALYZING THE RELATIVE CONTRIBUTION OF NIPA FACTORS

The basic NIPA model yields an estimate of the dollar contribution of each factor to the growth in net income. An example of the model output is shown in Table 2. However, these dollar estimates can be affected by a number of factors and thus could vary substantially from year to year. Furthermore, a change in the magnitude of a given factor from year to year is hard to interpret because of the continuously shifting base year. For example, the deflated quantities such as revenues in any two years are expressed in terms of the prices of the previous year. Thus the difference between the deflated revenues in two years within the basic NIPA framework cannot be treated merely as a change in the "real" or physical volume of business.

This difficulty can be avoided by comparing the percentage contribution of each of the factors relative to the subtotals for the "income-augmenting" and the "income absorbing" factors respectively (% f(AUG,i)) and % f(ABS,i)), computed as follows:

$$\% f(AUG,i) = Fi/TAUG \qquad i = 1, 2, 3, 4$$

$$\% f(ABS,i) = Fi/TABS \qquad i = 5, 6, \ldots, 14.$$

where Fi are the various NIPA factors and

TAUG = Subtotal for the Income-Augmenting Factors; and

TABS = Subtotal for the Income-Absorbing Factors.

The resulting percentage distribution is shown in Table 2.

While these percentage factors are a little more stable over time, compared with the dollar contributions, they are also subject to several influences whose importance can vary from year to year and which are not explicitly quantified in NIPA. Moreover, the two subtotals themselves are arbitrary and do not bear any direct relationship to any of the financial variables that financial planners have to work with. Thus we propose the following normalized NIPA factors, using the level of net income in the previous year as the normalizing variable.

$$Gi(t) = Fi(t)/NI(t-1)$$

where

$Fi(t)$, $i = 1, 2, . . ., 14$ are the dollar contributions of the NIPA factors in the current year; and

$NI(t-1)$ = Level of Net Income in the previous year.

By definition,

$$\sum_{i=1}^{14} Fi(t) = dNI(t).$$

Thus

$$\sum_{i=1}^{14} Gi(t) = dNI(t)/NI(t-1)$$

Gi(t) can be interpreted as the percentage contribution of the ith factor in year t to the growth rate of net income in that year. That is, each Gi(t) is a proportionate growth factor and that all of them combined account for the total growth in net income during the year.

This normalization procedure is appealing because,

(a) it uses the level of net income in the previous year as the normalizing variable, which is independent of the current year's distribution of NIPA factors themselves; and

(b) it directly shows the importance of a given factor in determining the growth rate of net income.

Table 3 shows a three-year history of these proportionate growth factors (along with projected results for 1980-1985 which will be discussed in Section IV) for the XYZ Corporation. For example, in 1979, productivity accounted for 32% of previous year's net income, earnings on capital 16%, rate changes 12% and other income 1%. The combined contribution of all income augmenting factors in that year was 61%. This means that if there had been no inflation in MR&S, Labor, etc., and no increases in taxes, depreciation or interest charges, etc., net income would have grown by 61%. But unfortunately, all of these factors were present. For example, inflation in MR&S amounted to 11% of previous year's net income and inflation in Labor costs (including Social Security taxes) another 30%. However, there was some relief from the Property and Other Non-income Taxes (-3%) and from Federal Income Taxes (-4%). The combined negative effect of the income absorbing factors was to reduce the net income growth by 55%. Hence the actual percent growth in net income of 6% (= 61-55) in 1979.

As shown in Table 3, the relative importance of productivity over the three year historical period is fairly stable around 30%. Capital expansion ranges between 16% and 18% while the contribution of price changes varies widely. The latter is a reflection of the irregular price adjustment process in this particular case where price changes come in lumps.

On the negative side, inflation in MR&S ranging between 8% and 11% is a direct result of the changes in the general price level for materials and services the firm buys from other firms. Similarly, inflation in labor costs broadly reflects the increases in wage rates and related benefits, as well as changes in the Social Security tax rates. This factor varies widely between 18% and 30% and it partly reflects the effects of a three-year bargaining cycle and changes in Social Security tax legislation. Depreciation is very stable around 12%, whereas income taxes and other financial factors show considerable variation.

It should be noted that we are not necessarily implying that the normalized NIPA factors should remain stable over time. But when a particular factor shows a significant change, it should be regarded as a signal of a fundamental shift that should be investigated further.

IV. ALTERNATIVE PLANNING SCENARIOS

We are now ready to investigate the behavior of net income under varying assumptions for the future, and to use NIPA to analyze how various factors contribute to changes in net income. In this section we present three illustrative cases. The first case assumes that labor productivity increases in the planning period at the same average rate as it did in the last five years (Simulation or Sim A); in the second we hold the absolute level of labor productivity constant, i.e., zero growth in labor productivity (Sim B); and in the third, it grows at a rate 20% faster than the average growth in the past five years (Sim C). All other variables for the three simulations are projected according to the assumptions sketched out in the table on the next page.

While it is possible to generate many other scenarios with a model like this, we have focused on the effect of varying productivity on the bottom line. Given certain assumptions about the behavior of productivity, we first derive a complete income statement. Then by solving the NIPA equation system, the resulting changes in net income are analyzed in terms of the NIPA factors described in the earlier sections.

In terms of the schematic in Figure 2, our key assumptions about labor productivity primarily affect the Productivity Module through impact on hours while other assumptions impact the Price Effects module. The Capital Growth module is affected only in Sim C where we make the additional assumption that the capital-labor ratio be held constant.

Assumptions for Some Alternative Future Scenarios for 1980-1985

Key Variables	Sim A	Sim B	Sim C
Output Volume	Average Rate of Growth	Average Rate of Growth	20% Above Average Growth in Volume
Output Prices	Average Rate of Increase	Average Rate of Increase	Average Rate of Increase
Employee Hours	Average Labor Productivity Gain	Zero Labor Productivity Gain	20% Above Average Labor Productivity Gain
Hourly Compensation	Average Rate of Increase	General Economy's Increase for Hourly Comp.	General Economy's Increase for Hourly Comp.
Capital Stock	Average Rate of Increase	Average Rate of Increase	Constant K/L Ratio at Average Level
Prices of Plant & Equipment	Average Rate of Increase	General Inflation in PDE Prices	General Inflation in PDE Prices
Materials	Average Rate of Increase	Avg. Vol. Growth, General Inflation in Prices	Avg. Vol. Growth, General Inflation in Prices
Depreciation	Average Rate of Increase	Average Rate of Increase	Average Rate of Increase
Interest	Average Rate of Increase	Average Rate of Increase	Average Rate of Increase
Income Taxes	Computed at the Avg. Tax Rate	Computed at the Avg. Tax Rate	Computed at the Avg. Tax Rate
Other Factors	Average Rate of Increase	Average Rate of Increase	Average Rate of Increase

This means that once the required hours have been determined, we must further determine the level of capital which is consistent with those hours. Note that while our assumptions were centered around labor productivity which directly affected the required hours (or labor input), NIPA still utilizes the total productivity concept in making all calculations shown in Tables 1 thru 3. By focusing on labor productivity, we are able to isolate the impact of this key factor on the financial performance of the firm. Of course, similar analyses could be conducted on any of the other variables of interest to the planners.

Before discussing the results of the three simulations, however, it might be useful to describe the process by which we have explicitly incorporated the labor productivity behavior into the income statement so that the distinction among the three simulations can be better understood.

Let the labor productivity ratio be

$$LP(0) = Q(0)/L(0)$$

where Q is total output and L is total hours in the base period, say 1979.

For Sim A, LP is assumed to grow at the average rate of increase(g) of the last 5 years. That is,

$$LP(t') = LP(0). (1 + g)^{t'}$$

where t' refers to the planning period only. Thus given an exogenously determined output Q(t'), the required hours become

$$L(t') = Q(t')/LP(t').$$

In Sim B, we set LP(t') = LP(0) for all planning periods and derive the hours that would be required to produce the output growing at the average rate of growth determined by supply and demand. That is, the required hours are

$$L(t') = Q(t')/LP(0)$$

Simulation C is somewhat more complicated by the fact that in addition to assuming that labor productivity grows at a rate 20% faster than the average for the last five years, we further assume that the capital labor ratio is constant during the planning period*. This means that physical capital requirements have to be estimated consistent with the two assumptions. Thus we first calculate

$$L(t') = Q(t')/LP(t')$$

and then compute capital K(t'), as

$$K(t') = L(t') \cdot \frac{K(0)}{L(0)}$$

Even though K(t') is not needed for deriving the income statement, it will be needed in running NIPA to compute the capital growth factor and productivity.

L(t') is used along with assumed increases in the hourly rate of compensation (including all labor related costs such as wages, overtime, benefits and employment taxes) to derive total labor costs for the respective simulations. Given revenues and all expenses, we first

* Alternatively, we could have assumed that the capital labor ratio continues to rise at the average rate of the last five years. However, our simpler assumption of a constant capital labor ratio measured in terms of unaugmented capital and labor inputs might be quite realistic if it is true that technological change improves the efficiency of capital to a greater degree than that of labor, so that in effect, capital in efficiency units per unit of labor is rising.

calculate gross income before taxes. For simplicity, we calculate Federal and State and Local income taxes by using the respective average effective tax rates in 1979. Finally, we compute the after-tax net income which then becomes the focus of the NIPA calculations.

Clearly, all of these simulations would present a more realistic picture if we were to use an econometric forecasting model to project demand, production and financial variables and then apply NIPA*. However, for this paper we have chosen to focus on the interpretation and in-depth analysis of the various planning scenarios rather than on the accuracy of projections. Hence the simplicity of our assumptions in generating these hypothetical scenarios.

As an example of the complete set of analytical results currently generated by NIPA, we show the following data in Tables 1 thru 3 for Sim A for the planning period 1980 to 1985, along with actual results for 1976 to 1979.

Table 1: NIPA Summary

Table 2: Percentage Distribution of NIPA Factors

Table 3: Relative Importance of NIPA Factors

* One such model is described by B.E. Davis, G. Caccapollo and M.A. Chaudry, "An Econometric Planning Model for American Telephone and Telegraph Company," Bell Journal of Economics, Spring 1973. Also see M. Werner, "Productivity Based Planning Model for Teleglobe Canada," Proceedings of the International Telecommunications Conference, 1979.

For comparative purposes, we have plotted the key variables from the three simulations in Figures 3 thru 7. For example, Figure 3 shows the three assumptions about labor productivity; Sim A with average growth in output per hour; Sim B with zero growth and Sim C showing 20% higher than the average growth rate. We see the dramatic impact of zero productivity on net income which takes a nose dive starting in 1980 and ending up negative in 1984 and 1985 (see Table 4). Figure 4 shows this phenomenon in terms of changes in net income, contrasting Sim B with the alternative scenarios. For instance under Sim B, the net income loss increases by nearly $4 billion in 1985 whereas A and C show positive gains.

If we had looked at the nominal net income data in Table 4 and the total revenue and total expenses alone, we would not be able to easily identify the source of the decline in net income. It could have resulted from any number of causes including higher wages, materials costs and other inflationary pressures. NIPA on the other hand, provides a vivid analysis of the true picture. With no growth in labor productivity, the number of employee hours required to produce the growing output increases rapidly resulting in substantial increases in the dollar value of labor input, which reduces the contribution of total factor productivity turning it into losses in 1983-1985. This is evident in Figure 5 showing the dollar contribution of TFP under the three scenarios. We can further examine the implications of zero labor productivity by looking at the dollar and relative impacts of inflation in labor (Figures 6 and 7 respectively). In both dollar and relative terms, we see that Sim B is substantially higher than either A or C, even though increases in the hourly compensation in B and C are the same. This difference occurs because the increase in the hours in B is so much larger than in C (where productivity rises 20% faster than the average of the last 5 years). The relative importance of the underlying productivity growth is especially highlighted in Figure 7 which shows the resulting increase in labor costs (due to the the increases in hourly compensation coupled with labor productivity behavior) as a percent of previous year's net income.

V. CONCLUSIONS

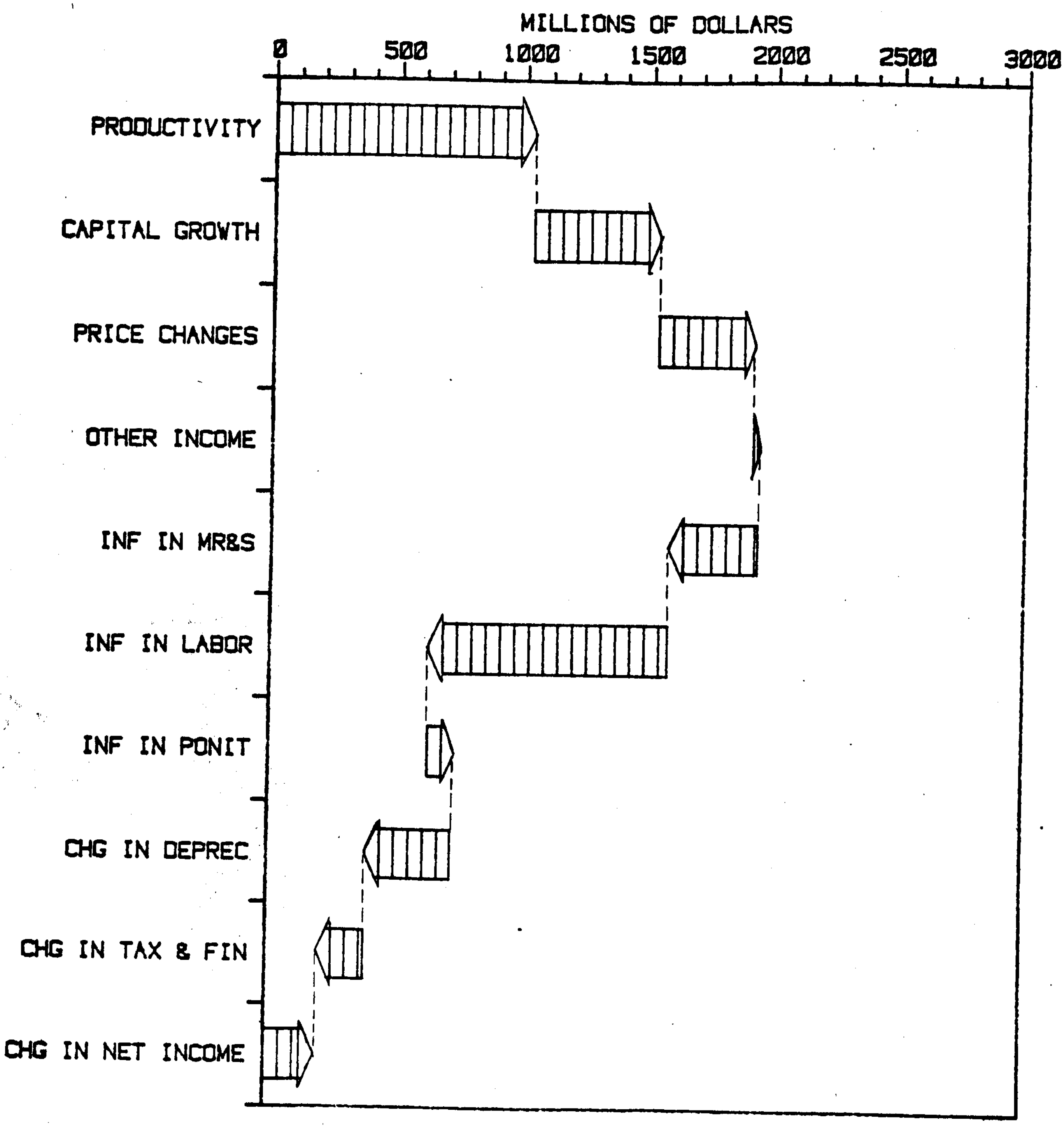
We have described a planning model which provides a great deal more information than most financial models offer. NIPA permits the user to account for productivity and many other underlying factors explicitly and in terms of dollars and cents as they affect the bottom line. Because of this feature, managers find NIPA easier to understand than most models offered to planners. Thus they are more inclined to use it as a planning tool and be able to put the dollar productivity estimate in proper perspective of the income statement. It is worth noting that most managers have shyed away from the use of traditional productivity measures expressed in terms of percentage growth rates, because they could not use such data in any direct way. The most some planners were able to do was to compare the projected productivity growth with the past record and made a qualitative judgement as to whether the budget based on that projection was a reasonable one. The only control these managers could exercise was to demand an explanation of the poor productivity built into the budget and ask the operating entity anticipating lower than averge productivity to redo its budget with some target productivity growth. With NIPA, the planner is able to see the dollar contribution of productivity to growth of the bottom line and is thus able to set a specific quantitative target the entity must achieve if it is to meet its budget goal. Moreover, looking at the standard NIPA Summary results, the planner is able to see whether it is poorer output growth, faster input growth or worse inflation beyond the control of the managers which is the culprit. In other words, while NIPA cannot come up with solutions, it can at least point out the problem areas which should be investigated in search of management options to intervene. Moreover, it provides enough disaggregation of the underlying factors to allow a distinction to be made between what management can and cannot control and thus act accordingly.

NIPA is currently operational in an interactive mode and allows the user to provide a variety of inputs and exercise many options in terms of generating alternative scenarios and select desired results. However, this means that the necessary budget data must be prepared in advance through whatever budgeting process may be in use. We plan to extend the model to include target setting by the user and having the model solve for all the endogenous varibles before making the standard NIPA calculations. For instance, we could set a target for net income or rate of return and then given specific operating rules, demand conditions and appropriate constraints, solve for the necessary inputs to NIPA.*

This would permit the user to construct a budget, analyze it and alter it with the help of NIPA to achieve prespecified management goals. For example, such a model could also be used for determining the change in hours or other resources that would be required if a certain change in the budgeted net income is to be made.

* A similar model has been proposed by M. Werner (1979), using a somewhat different alternative rationale.

Figure 1

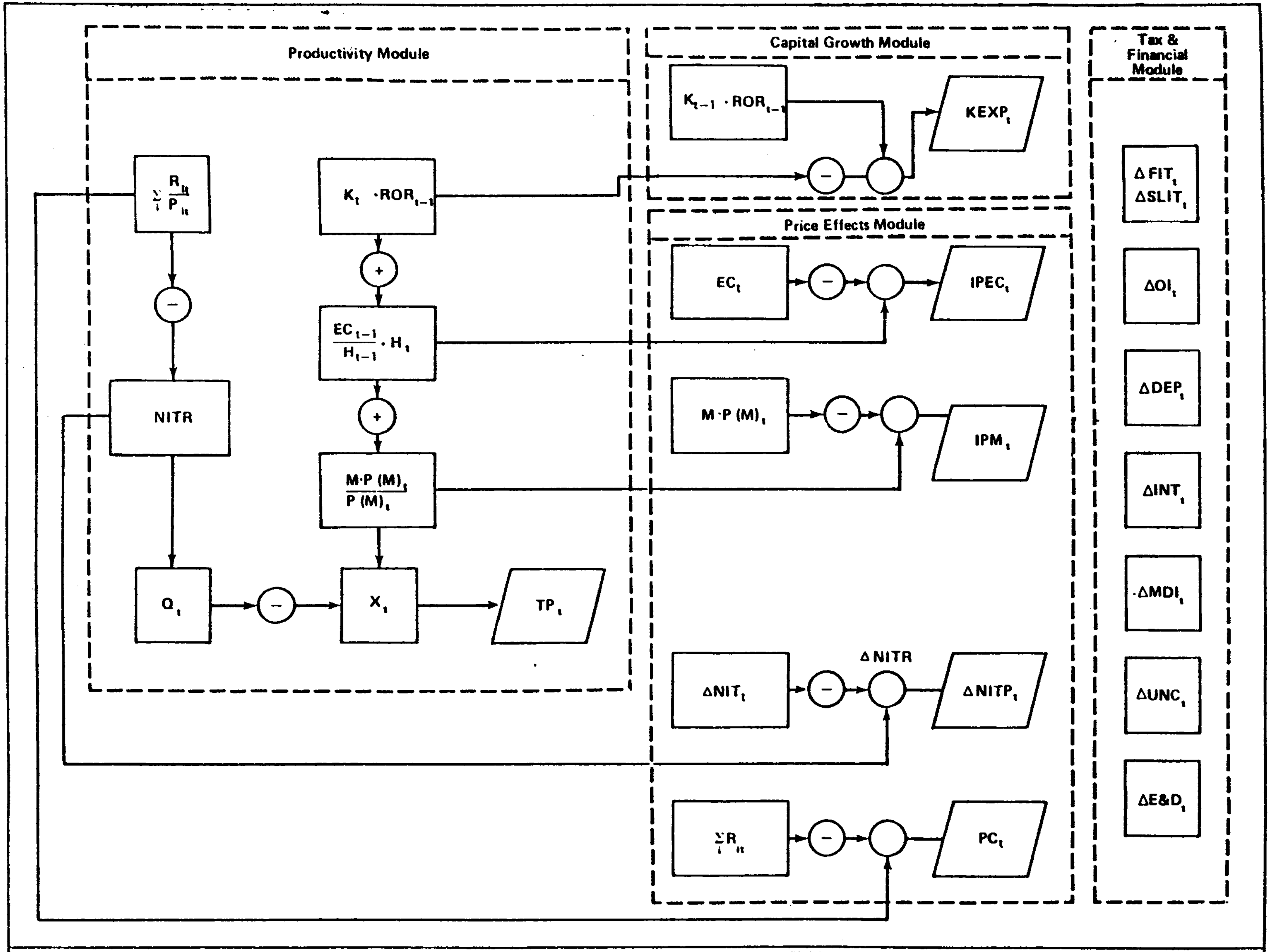


FACTORS AFFECTING CHANGE IN NET INCOME

XYZ Corporation - 1979

Figure 2

NET INCOME AND PRODUCTIVITY ANALYSIS (NIPA) MODEL



$\Delta NINC = TP + KEXP + PC + \Delta OI - IPM - IPEC - \Delta NITP - \Delta FIT - \Delta SLIT - \Delta DEP - \Delta INT - \Delta MDI - \Delta UNC + \Delta E\&D$

Note: H = L = Hours.

Figure 3

LABOR PRODUCTIVITY

UNDER THREE ALTERNATIVE SCENARIOS

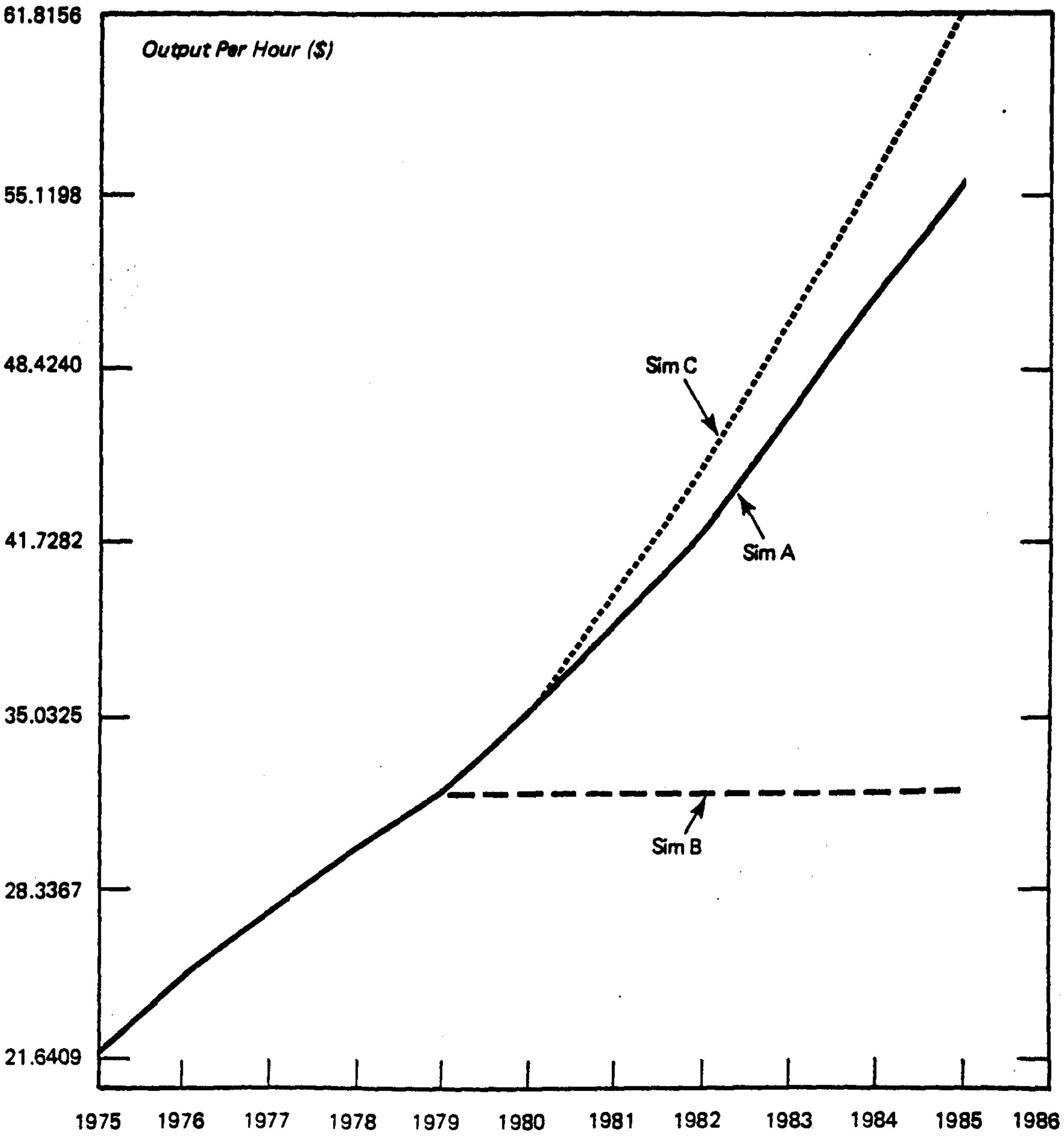


Figure 4

CHANGE IN NET INCOME

UNDER THREE ALTERNATIVE SCENARIOS

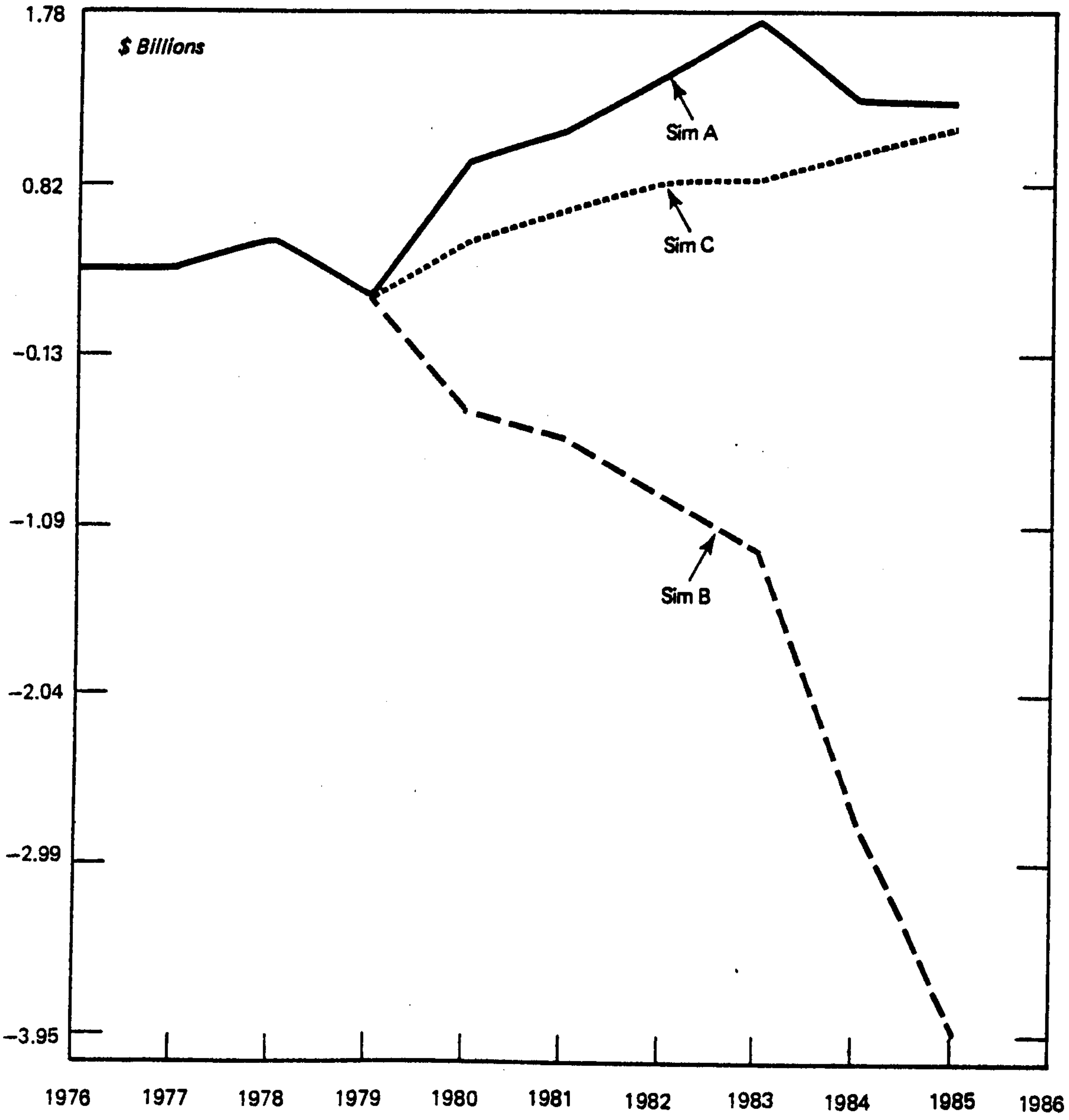


Figure 5

TOTAL FACTOR PRODUCTIVITY CONTRIBUTION

UNDER THREE ALTERNATIVE SCENARIOS

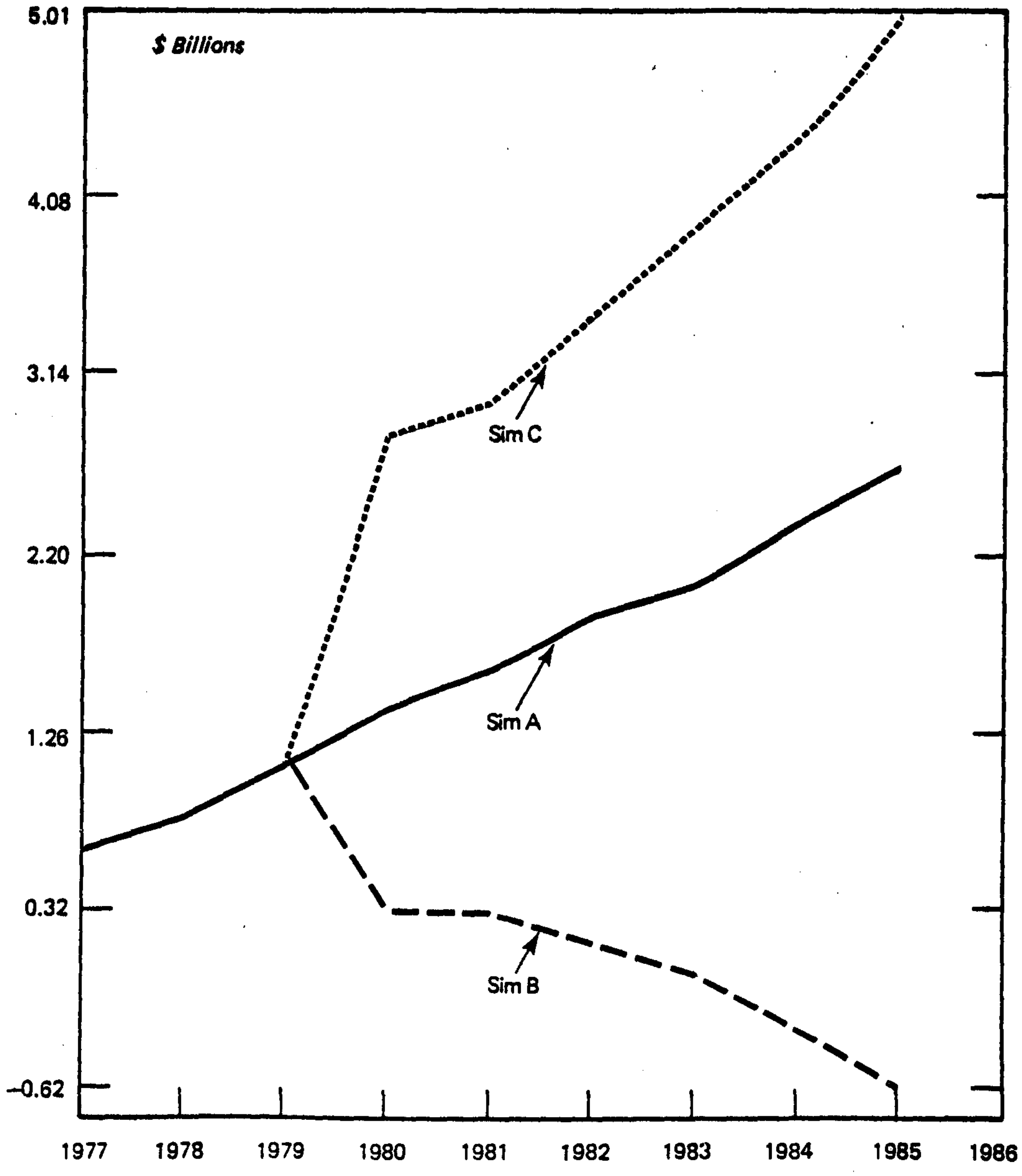


Figure 6

DOLLAR IMPACT OF INFLATION IN LABOR

UNDER THREE ALTERNATIVE SCENARIOS

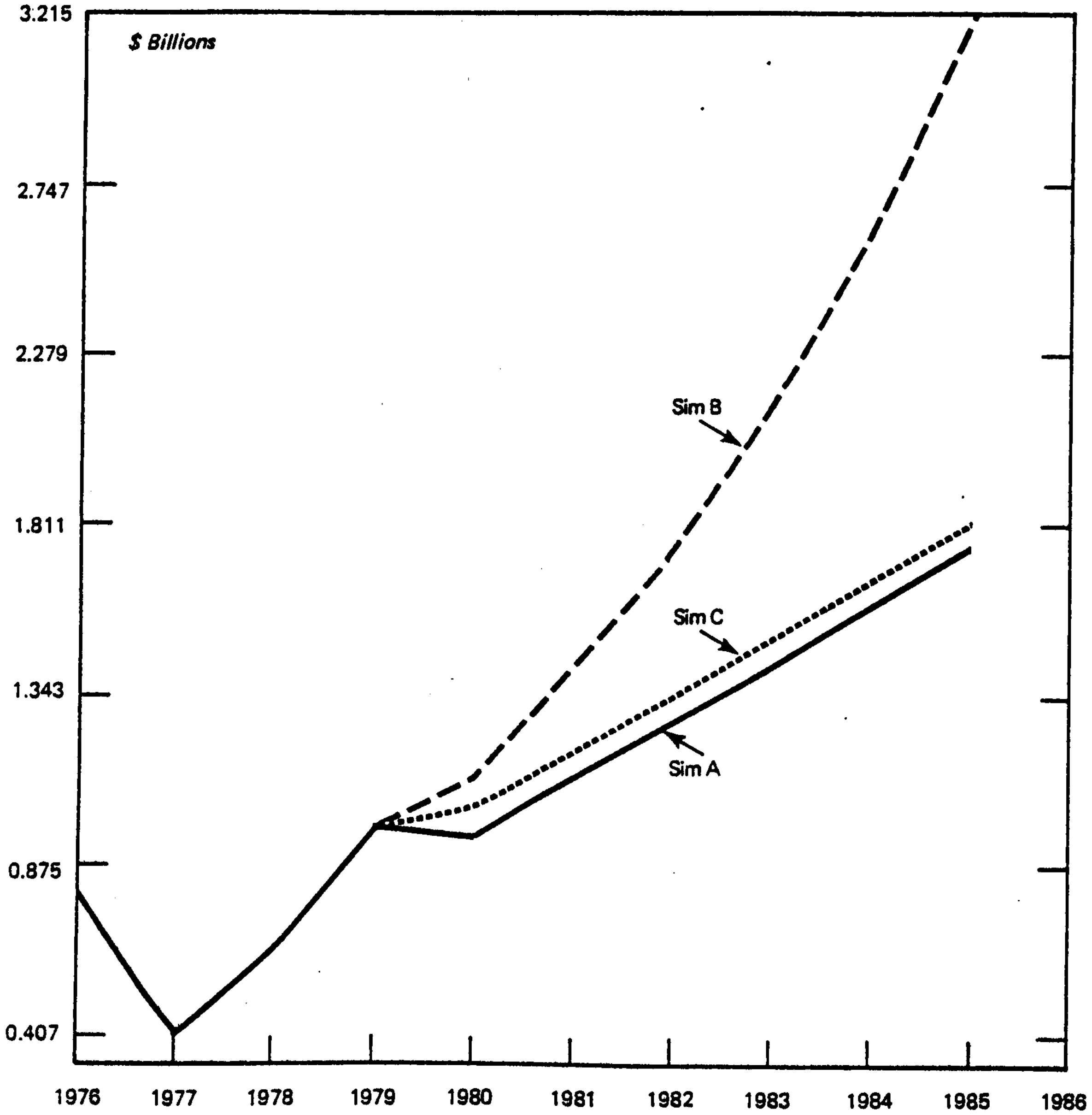


Figure 7

RELATIVE IMPACT OF INFLATION IN LABOR

(% OF NET INCOME IN THE PREVIOUS YEAR)

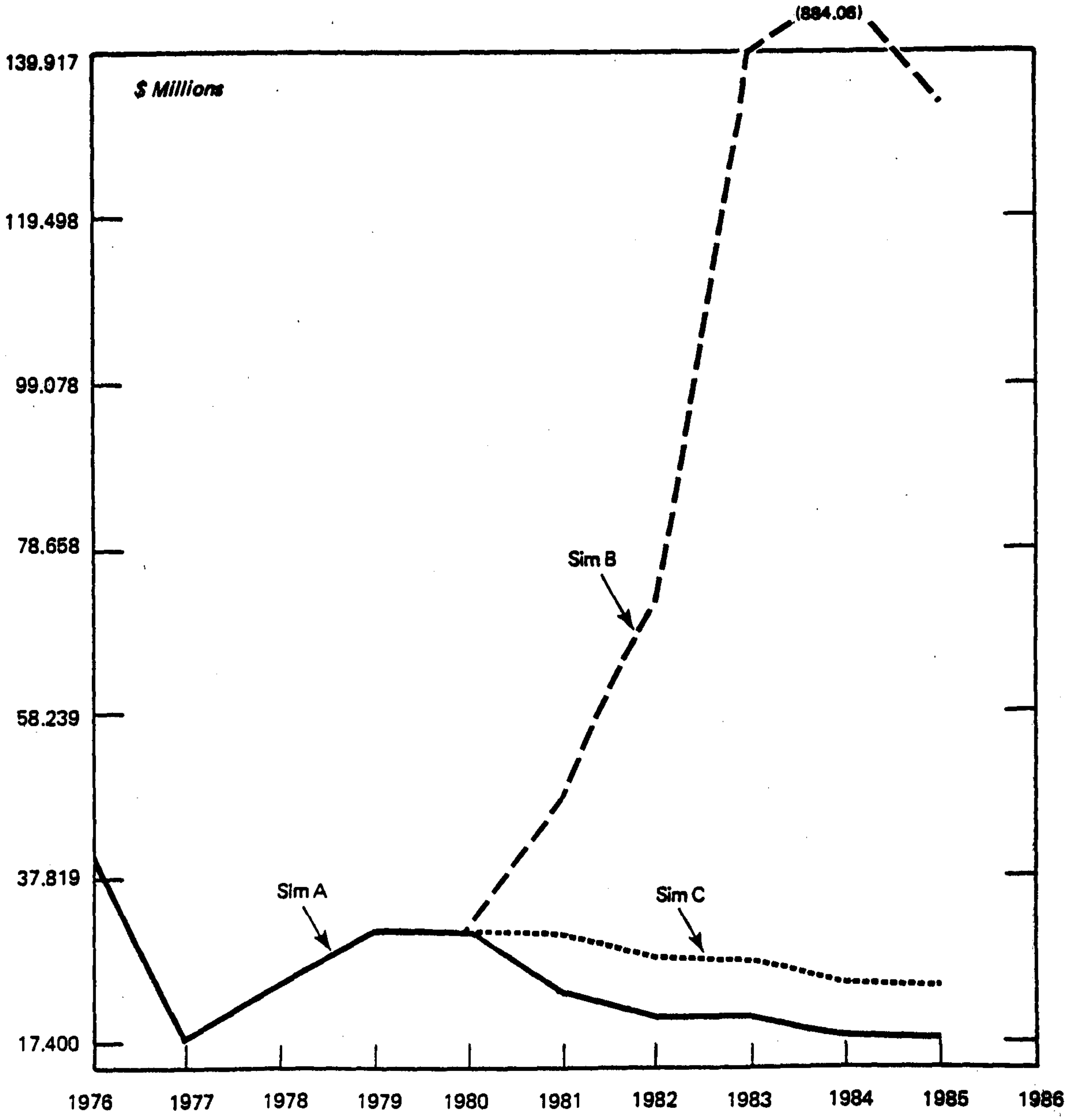


Table 1

XYZ CORP - AVG GROWTH IN O/H
NET INCOME AND PRODUCTIVITY ANALYSIS (NIPA)
(MILLIONS OF DOLLARS)

YEARS IN STUDY	1977	1978	1979	1980	1981	1982	1983	1984	1985
CHG IN DEFL REVENUES	$1,565.8	$2,060.5	$2,223.1	$2,157.2	$2,427.0	$2,732.1	$3,076.8	$3,466.5	$3,907.4
- CHG IN DEFL PONIT	79.3	98.2	93.8	78.0	64.1	69.8	76.0	82.8	90.3
= CHG IN OUTPUT	1,486.5	1,962.3	2,129.3	2,079.2	2,362.9	2,662.3	3,000.8	3,383.7	3,817.1
- CHG IN CAPITAL INPUT	382.4	496.5	504.8	407.9	239.9	268.1	299.7	335.0	374.6
- CHG IN LABOR INPUT	328.0	378.6	351.0	54.2	250.3	276.4	314.3	357.8	407.6
- CHG IN MRAS INPUT	215.5	264.0	247.8	187.3	240.0	275.4	316.1	362.9	416.7
= PRODUCTIVITY GAIN	560.6	823.2	1,025.7	1,429.8	1,632.7	1,842.5	2,070.8	2,328.1	2,618.2
EARNINGS ON CAPITAL EXPANSION	382.4	496.5	504.8	407.9	239.9	268.1	299.7	335.0	374.6
RATE CHANGES	568.2	619.8	379.8	994.0	1,115.9	1,250.6	1,402.7	1,574.4	1,768.6
OTHER INCOME	159.0	84.2	18.2	64.4	71.4	79.2	87.9	97.5	108.2
TOTAL POSITIVE FACTORS	1,670.2	2,023.7	1,928.4	2,896.0	3,059.9	3,440.4	3,861.0	4,335.0	4,869.5
INFL IN MR&S	184.3	245.8	355.5	447.4	488.5	560.7	643.5	738.4	847.3
INFL IN LABOR INCL SS TAXES	407.2	607.8	943.5	989.5	1,100.5	1,238.8	1,391.5	1,559.3	1,743.5
INFL IN PONIT	41.1	-36.7	-96.2	30.4	52.6	56.0	59.7	63.5	67.7
CHG IN DEP DUE TO INFLATION	98.0	82.5	170.2	221.3	320.4	354.6	392.4	434.3	460.6
CHG IN DEP DUE TO OTHER EFFECTS	224.1	201.9	171.2	155.4	96.4	106.7	118.0	130.6	144.5
CHG IN FEDERAL INCOME TAXES	180.3	307.5	-142.4	-100.5	-379.6	-539.5	-732.6	-140.2	-
CHG IN STATE & LOCAL INCOME TAXES	35.0	20.4	11.6	-33.1	-37.1	-52.8	-71.6	-13.7	-
CHG IN INT EXP DUE TO INT RATES	-10.6	41.4	136.6	29.6	64.2	69.7	75.7	82.1	89.2
CHG IN INT EXP DUE TO DEBT VOLUME	34.7	68.8	120.4	122.5	100.9	109.5	118.8	129.0	140.0
CHG IN UNCOLLECTIBLES	20.5	46.4	67.9	52.3	62.6	74.9	89.6	107.2	128.3
CHG IN MISC DEDUC FROM INCOME	-	-	-	-	-	-	-	-	-
LESS:CHG IN EXTRA & DEL ITEMS-NET	-31.3	41.0	6.7	-1.3	-.4	-.1	-.0	-.0	-.0
TOTAL NEGATIVE FACTORS	1,246.0	1,544.8	1,731.5	1,916.0	1,869.8	1,978.7	2,085.0	3,090.7	3,641.1
ESTIMATED CHG IN NET INCOME	424.2	478.9	196.9	980.1	1,190.2	1,461.7	1,776.0	1,244.3	1,228.5
ACTUAL CHG IN NET INCOME	424.2	478.9	196.9	980.1	1,190.2	1,461.7	1,776.0	1,244.3	1,228.5

Table 2

XYZ CORP - AVG GROWTH IN O/H
NET INCOME AND PRODUCTIVITY ANALYSIS (NIPA)
(PERCENTAGE DISTRIBUTION)

YEARS IN STUDY	1977	1978	1979	1980	1981	1982	1983	1984	1985
PRODUCTIVITY GAIN	34	41	53	49	53	54	54	54	54
EARNINGS ON CAPITAL EXPANSION	23	25	26	14	8	8	8	8	8
RATE CHANGES	34	31	20	34	36	36	36	36	36
OTHER INCOME	10	4	1	2	2	2	2	2	2
TOTAL POSITIVE FACTORS	100	100	100	100	100	100	100	100	100
INFLATION IN HR&S	15	16	21	23	26	28	31	24	23
INFLATION IN LABOR INCL. S.S. TAXES	33	39	54	52	59	63	67	50	48
INFLATION IN PONIT	3	-2	-6	2	3	3	3	2	2
CHG IN DEPRECIATION EXPENSE	26	18	20	20	22	23	24	18	17
CHG IN FEDERAL INCOME TAXES	14	20	-8	-5	-20	-27	-35	-5	-
CHG IN STATE & LOCAL INCOME TAXES	3	1	1	-2	-2	-3	-3	-	-
CHG IN INTEREST CHARGES	2	7	15	8	9	9	9	7	6
CHG IN UNCOLLECTIBLES	2	3	4	3	3	4	4	3	4
CHG IN MISC DEDUC FROM INCOME	-	-	-	-	-	-	-	-	-
LESS:CHG IN EXTRA & DEL ITEMS-NET	-3	3	-	-	-	-	-	-	-
TOTAL NEGATIVE FACTORS	100	100	100	100	100	100	100	100	100

Table 3

XYZ CORP - AVG GROWTH IN O/H
RELATIVE IMPORTANCE OF NIPA FACTORS
(PERCENT OF PREVIOUS YEAR'S NET INCOME)

YEARS IN STUDY	1977	1978	1979	1980	1981	1982	1983	1984	1985
PRODUCTIVITY GAIN	25	31	32	42	38	33	30	27	26
EARNINGS ON CAPITAL EXPANSION	17	18	10	12	6	5	4	4	4
RATE CHANGES	25	23	12	30	26	23	20	18	16
OTHER INCOME	7	3	1	2	2	1	1	1	1
% CHG IN TOTAL POSITIVE FACTORS	74	75	61	86	70	62	55	49	49
INFLATION IN MR&S	8	9	11	13	11	10	9	8	8
INFLATION IN LABOR INCL. S.S. TAXES	18	23	30	29	25	22	20	18	17
INFLATION IN PONIT	2	-1	-3	1	1	1	1	1	1
CHG IN DEPRECIATION EXPENSE	14	11	11	11	10	8	7	6	6
CHG IN FEDERAL INCOME TAXES	8	11	-4	-3	-9	-10	-10	-2	-
CHG IN STATE & LOCAL INCOME TAXES	2	1	-	-1	-1	-1	-1	-	-
CHG IN INTEREST CHARGES	1	4	8	5	4	3	3	2	2
CHG IN UNCOLLECTIBLES	-	-	-	-	-	-	-	-	-
CHG IN MISC DEDUC FROM INCOME	1	2	2	2	1	1	1	1	1
LESS:CHG IN EXTRA & DEL ITEMS-NET	-1	2	-	-	-	-	-	-	-
% CHG IN TOTAL NEGATIVE FACTORS	55	57	55	57	43	36	30	35	36
EST % GROWTH IN NET INCOME	19	18	6	29	27	26	25	14	12
ACTUAL % CHG IN NET INCOME	19	18	6	29	27	26	25	14	12

Table 4

ALTERNATIVE NET INCOME SCENARIOS
(Millions of Dollars)

	Sim A	Sim B	Sim C
1976	2,268	-	-
1977	2,692	-	-
1978	3,171	-	-
1979	3,368	-	-
1980	4,348	3,244	3,940
1981	5,538	2,934	4,596
1982	7,000	2,405	5,370
1983	8,776	1,562	6,269
1984	10,020	-613	7,322
1985	11,249	-3,818	8,533

Notes on key assumptions:

Sim A: For 1980-1985, all variables are assumed to grow at the average rate for the period 1974-1979.

Sim B: Assumes zero growth in output per hour (labor productivity) for 1980-1985.

Sim C: Output and output per hour are assumed to grow at a 20% faster rate but capital is derived by holding the capital labor ratio constant as of 1979.

Appendix

DEFINITIONS OF SYMBOL IN FIGURE 2

NI = Net Income in a Year

R = Total Revenue

R_i = Revenue from the ith Product

P_i = Price Index (deflator) for the ith Product

Q = Total Output

C = Total Cost

PC = The Value of Output Price Changes from the Previous Year

X_j = Quantity of jth input

P_j = Price of the jth input

M = Purchased Materials and Services (Deflated)

P(M) = Implicit Deflator for Purchased Materials and Services

IPM = Effect of Inflation in Materials Prices

EC = Total Employee Compensation, including Social Security taxes

IPEC = Effect of the Change in Labor Input Prices

H = Total Employee Hours (= L used in Section II)

w = Effective Hourly Rate of Remuneration
= EC/H

TP = Total Productivity Gain

GPI(K) = Gross Capital Price Index

PROPT = Property Taxes

GRT = Gross Receipts Taxes

CST = Capital Stock Taxes

ONIT = Other Non-Income Taxes

NIT = Total Non-Income Taxes
= TPROP+TGR+TCS+TONI

NITR = Real Component of Non-Income Taxes at the Previous Year's Tax Rate

NITP = "Price" (Tax Rate Change) Component of Non-Income Taxes

KGPE = Gross Invested Capital in Plant and Equipment
KAC = Average Cash Component of Capital
KNR = Average Net Receivables Component of Capital
KMS = Average Materials and Supplies Component of Capital
GK = Average Total Capital for the Year (=KGPE+KAC+KNR+KMS)
K = Deflated Total Capital (= GK/GPI(K))
ROR = Rate of Return on Total Capital for the Firm (Actual)
= P(K) as used in Section II
KEXP = Earnings on Capital Expansion
FIT = Federal Income Tax
SLIT = State and Local Income Taxes
INT = Total Fixed (Interest) Charges
DEP = Depreciation Expense (Book)
OI = Other Income
MDI = Miscellaneous Deductions from Income
E&D = Extraordinary and Delayed Charges & Credits - Net
UNC = Uncollectible Revenues

TOTAL FACTOR PRODUCTIVITY FOR MANAGEMENT:

THE POST-MORTEM AND PLANNING FRAMEWORKS

MICHAEL DENNY

Institute for Policy Analysis
University of Toronto

ALAIN DE FONTENAY

Government of Canada
Department of Communications

MANUEL WERNER

Telecommunications Consultant

I. Introduction

The question in most firms is not whether there is any preoccupation with productivity but rather the level at and the degree to which it is applied. There has, over the past 60 years, been considerable effort in the direction of first measuring, then improving, and ultimately, monitoring productivity. The sequence is probably repeated to different levels of sophistication in most departments or areas of activity. It is certainly highly pervasive within the actual operating areas, such as the plant floor, work sites and so on. Briefly an inquiry about productivity, in almost any firm, would not be met by a blank stare. However, inquiring as to the significance of these micro-applications of partial productivity measures to overall corporate performance would almost certainly not elicit an informed response. Further, inquiring about the role of all the micro-measures in the corporate planning exercise would elicit even less of a response. Given the partial nature of all the diverse productivity and quasi-productivity measures in use at the detailed activity levels of the firm, it would be almost impossible to make any meaningful connection with some global type of measure. This is not to imply that these micro-measures are in some sense unimportant when, on the contrary, they are probably an excellent cost control tool for section, division or department managers. The only point of contention lies rather with the inability to string them together for ultimate use in corporate/budgetary planning. To draw together the diverse inputs and outputs of any large firm requires a somewhat more global measure of productivity. The theme of this paper centres on the analytical and planning models that are integrated into the planning process on the basis of a Total Factor Productivity (TFP) measure.

A series of related productivity models for management will be introduced. We will start with the by-now standard NIPA (Net Income Productivity Analysis) model which is a purely descriptive and passive management tool, and then present the UNIPA (Unconstrained NIPA) model which enables the firm to compare its rate of return to the capital market. Third we will show how the UNIPA model can be used for a post-mortem analysis, through which the firm can evaluate its success in meeting its planned budget. Finally we will introduce the PAP (Productivity Analysis for Planning) model which the firm can use as a top-down guideline and control in its corporate budgeting and planning.

The first, NIPA, is a model developed to analyse the impact of productivity both historically and within the context of a fully developed financial plan. It is oriented, in particular, towards explaining the growth in Net Income, which, for the management of the firm, is the most important single statistic which they monitor. For them, it is the complex which most clearly mirrors performance. It is for this reason that the productivity model has been designed around Net Income growth as its reference point. From a purely economic perspective there is nothing unnatural about this approach. While the accountant views Net Income as a residual return to invested capital, the economist sees it as both a cost of and a return to invested capital. Considering net income, as the value of some quantity of capital that is supplied to the enterprise at a fixed price per unit, along with the trade-off between quantity and productivity (as dictated by price movements) the model will be seen as just another more elaborate view of the basic profit statement. In that it is merely a decomposition and subsequent rearrangement of the basic price and quantity income statement components, many different presentations of the same data are possible. Clearly, each of the various presentations will emphasize different aspects. We will examine them below. We will begin by a summary and brief commentary of the version developed independently at AT&T and Teleglobe Canada.[1] Following that overview of the NIPA, we will introduce, as a tool to compare the firm's earnings performance to what it can expect on the capital market, the UNIPA model. Fundamentally, we will remove the identity between revenue and cost which in the NIPA analysis is used to define residually the cost of capital, and we will allow for profits or losses.

In particular, our version resembles a combination of the analytical models at Electricité de France (Reimeringer (1980)) which do not constrain the return to capital to always equal its cost, and NIPA, which does not admit the possibility that planned and actual costs and revenues may not always be equal. Finally, we shall show how the UNIPA models can be used as a post-mortem and quasi-planning model. It can analyse historical performance and, as well, review future plans with a view to identifying the implicit productivity gains (or losses) and their impact on Net Income growth. In their present form NIPA and UNIPA models do not, in contrast to the PAP model also presented below, actually generate the plan.

The PAP is a pure planning model designed to develop a complete budgetary/corporate plan, at a fairly aggregate level, where the components of the various financial/accounting summaries all embody certain key management and corporate targets. More succinctly we may view this as something of a pure or guideline theoretical budget generated for top management so that they can more intelligently guide the longer more tedious development of a full-blown, bottom-up corporate budgetary plan. With the results of the planning model the process becomes far less arbitrary. The planners are in a position to prescribe unique upper limits for all the key financial statement items including labour and other expenses and the size of the capital budget. They are armed with the knowledge that any overshooting of these benchmark expense and expenditure figures will ensure that some or all of the preset targets will not be attained. While there are a whole array of possible targets, our model is built around what we believe to be the most important of these: the required return to invested capital (r), the forecast demand for the firm's production and, the desired growth in productivity.

II. NIPA

a) Introduction

Productivity gains or losses play an essential role in the degree to which a firm will succeed. It is productivity that allows the firm to weather the ravages of input price inflation without resorting to excessive output price increases which could damage market share in a competitive environment or not be permissible by the regulator and thus harm capital market operations. Although these are facts acknowledged by any entrepreneur, there are not many who, if they even measure it, effectively tie productivity information into the overall management of the firm. It is unfortunate because, once the measurement problem (which is probably the major stumbling block) is resolved, productivity results can be integrated directly into a quasi-financial accounting framework for use by decision-makers. This is apparent when we look at the basic accounting identity.

$$\text{Revenues} \equiv \text{Costs}$$

where costs account for all payments including those required capital payments such as interest, taxes and return to equity holders. By looking directly at the price and quantity components, the accounting identity becomes

$$\begin{aligned} &(\text{Price of Outputs}) \times (\text{Quantity of Outputs}) \\ &\quad = (\text{Price of Inputs}) \times (\text{Quantity of Inputs}) \end{aligned}$$

and with the definition:

$$\text{Total Factor Productivity} = \frac{\text{Quantity of Outputs}}{\text{Quantity of Inputs}}$$

it follows that:

$$(\text{Price of Outputs}) = (\text{Price of Inputs}) \div \text{Total Factor Productivity}$$

In other words, the basic rule, embedded in the accounting identity says that the price of output should be such as to cover that part of the price of inputs which is not offset by gains in Total Factor Productivity (TFP). Although this is somewhat of an oversimplification, it nevertheless demonstrates the essential role of TFP, as an offset (either partial or complete) to input price inflation. While most firms will try to price at what they believe the market will bear, and thus maximize the residual, net income, the basic pricing rule embedded in the accounting identity does provide an excellent guideline for any market situation, including the regulated sector.

Accepting the premise of the pricing rule is not very difficult. The major source of inhibition lies rather with the practical aspects of implementation. These include (i) the index number problem; (ii) the data definition problem and (iii) the difficulty of relating the individual price and quantity elements of Revenues and Costs directly to a management decision designed to affect the bottom line of the firm's income statement. While issues (i) and (ii) are of paramount importance, they are given extensive treatment elsewhere in Denny, de Fontenay and Werner (1980) and de Fontenay (1980) and will be assumed away leaving us to deal only with the last difficulty. Given that economic theory already provides a very extensive coverage of this aspect, with pricing and production rules for any number of market/optimization-objectives combinations, it may seem redundant to write yet another on the subject. However, while economic theory may tell the entrepreneur what level of output should be produced and at which price it should be sold, given his production function, cost relationships and market organization, it does not provide any link with the realities of his income statement, balance sheet or funds flow statistics. In this paper we propose to do just that. Section I will examine current applied work at AT&T and Teleglobe Canada. Specifically it will look at the Net Income Productivity Analysis (NIPA) model, a version of which is also presently in use at Electricité de France. In addition a more powerful version of NIPA will be presented in the second part of Section 1. As an extension of these purely post-mortem quasi-planning type of management TFP models, Section II introduces a pure planning model with explicit consideration given to targeted productivity and financial variables. It is partially based on work by Werner (1979) for Teleglobe Canada.

b) The Model

The final question, after deriving the basic NIPA relationship, concerns the best approximation to its set of continuous variables. As part of the development of this management tool it will be useful to examine two approaches: (1) beginning with TFP growth as the difference between the logarithmic differentials of output and input it will be seen that the final discrete approximation is arbitrary and (2) by developing the NIPA statement through the application of Diewert's Quadratic Lemma (Diewert (1976)) we show that the final discrete accounting statement is exactly derived.

The traditional NIPA assumptions are based on product exhaustion and factor prices equal to the value of their marginal products. If, in addition, revenues are equal to costs in every period, where costs include a required return to invested capital then this implies that the entire process is characterized by constant returns to scale. Thus, given the definition of Total Factor Productivity

$$\dot{TFP} \equiv \dot{Q} - \dot{X} \tag{1}$$

where $Q = Q(q_1,\dots,q_n|p_1,\dots,p_n)$ is an index of output

$X = X(x_1,\dots,x_m|w_1,\dots,w_m)$ is an index of input where q_i and p_i denote respectively the quantity and price of the i-th output and x_j and w_j that of the j-th input. A dot over the symbol indicates a logarithmic differential (i.e., a proportional rate of change). The above assumptions state that if and only if,

$$\dot{R} \equiv \dot{C} \quad \text{then} \quad \dot{TFP} = \dot{W} - \dot{P} \qquad (2)$$

$$\text{where} \quad \dot{R} = \dot{P} + \dot{Q} \qquad \text{and} \qquad \dot{C} = \dot{W} + \dot{X}$$

and W and P are price indices of input and output, either of Q and P and of X and W being implicit, respectively.

Combining (1) and (2) we have

$$\dot{Q} - \dot{X} = \dot{W} - \dot{P} \qquad (3)$$

which is the point of departure of the standard NIPA model. Each of the terms is a weighted aggregate where

$$\dot{Q} = \sum_i s_i\dot{q}_i \; ; \quad \dot{X} = \sum_j \sigma_j\dot{x}_j \; ; \quad \dot{W} = \sum_j \sigma_j\dot{w}_j \; ; \quad \dot{P} = \sum_i s_i\dot{p}_i$$

If we let the x_j, $j = 1$ to 3 represent K, L and M respectively and w_j, $j = 1$ to 3 represent the prices of K, L and M, denoted as r, w and m, respectively, then (3) can be rewritten as

$$\dot{TFP}_t + \dot{P}_t = \sigma_{w_t}\dot{w}_t + \sigma_{m_t}\dot{m}_t + \sigma_{r_t}\dot{r}_t \qquad (4)$$

which tells us that the changes in input prices will be exactly offset, in any period by some combination of TFP gains and output price changes. By adding $\sigma_{r_t}\dot{K}_t$, a term commonly referred to as "Capital Growth", to both sides of (4) we have the new expression

$$\dot{TFP}_t + \dot{P}_t + \sigma_{r_t}\dot{K}_t = (\sigma_{w_t}\dot{w}_t + \sigma_{m_t}\dot{m}_t) + (\sigma_{r_t} \, d\ln r_t k_t) \qquad (5)$$

The last term on the RHS, σ_{r_t} dln rK , is the proportional change in capital costs which is composed of a price and quantity component, $\sigma_{r_t}\dot{r}_t$ and $\sigma_{r_t}\dot{K}_t$ respectively. The Capital Growth term $\sigma_{r_t}\dot{K}_t$ can be more easily understood by noting that if TFP, $\dot{P}_t$ $(\sigma_{w_t}\dot{w}_t + \sigma_{m_t}\dot{m}_t)$ were all zero and if the firm could expand its capital stock while maintaining the same rate of return on that stock, then it would then be able to increase its net income by the same proportion. The components of dln rK are changes in depreciated expenses, debt service costs, taxes and the return to invested capital. For each of the components, we can define ex post ratios r_ℓ , $\ell = 1,..,4$ as the ratio of the particular expense to the total stock of capital such that

$$rK = \sum_{\ell=1}^{4} r_\ell K$$

Then

$$d\ln r_t K_t = \sum_{\ell=1}^{4} \varepsilon_{\ell,t} \, d\ln r_{\ell,t} K_t$$

where $\varepsilon_{\ell,t} = \frac{r_{\ell,t}}{r_t}$ is the share of each of the four components of the capital cost to total capital cost. We may now rewrite (5) as

$$\dot{TFP} + \dot{P}_t + \sigma_{r_t}K_t = [\sigma_{w_t}\dot{w}_t + \sigma_{m_t}\dot{m}_t] + \sigma_{r_t}\varepsilon_{1_t}(d\ln r_{1,t}K_t) + \sigma_{r_t}\varepsilon_{2t}(d\ln r_{2,t}K_t)$$

$$+ \sigma_{r_t}\varepsilon_{3t}(d\ln r_{3t},K_t) + \sigma_{r_t}\varepsilon_{4t}(d\ln r_{4,t}K_t) \qquad (6)$$

The elements on the RHS of (6) are all identifiable components of the standard income statement (in terms of proportional changes), weighted by their share of the total cost, they represent:

$\dot{w}_t$ and $\dot{m}_t$	=	the price movements of labour and other operating expenses
$d\ln r_{1,t}K_t$	=	depreciation expenses
$d\ln r_{2,t}K_t$	=	debt service and other financial instrument expenses
$d\ln r_{3,t}K_t$	=	relevant tax expenses
$d\ln r_{4,t}K_t$	=	net income

Expression (6) is nothing more than a decomposition of the basic accounting identify,

$$\dot{NI} = \dot{R} - \dot{C}^*$$

$$\text{where} \quad \dot{C}^* = \dot{C} - \dot{NI} \quad ; \quad \text{i.e.} \quad \dot{NI} \ \text{includes all capital costs}$$

The discrete approximation of (6) takes account of the facts that

(a) $\dot{z} = d\ln z = \frac{dz}{z}$; for z representing any of the dotted variables

(b) $\sigma_{z_t} = \frac{P_z z}{R}$; for P_z representing any input price

(c) $\varepsilon_{\ell,t} = \frac{r_{\ell,t}K_t}{r_t K_t}$ and $\sigma_{r,t}\varepsilon_{\ell,t} = \frac{r_{\ell,t}K_t}{R_t}$ where $\ell = 1,\ldots,4$

(d) $\dot{P}_t = \frac{PQ}{R}\frac{dP}{P}$

(e) $\dot{TFP} = \frac{1}{R}(PQdQ - WXdX)$.

Multiplying (6) by R and cancelling all the common terms leaves,

$$dTFP + QdP + rdK = [Ldw + Mdm] + [Kd\delta + \delta dK] + [Kd\phi + \phi dK]$$
$$+ [Kd\theta + \theta dK] + [Kd\pi + \pi dK] \qquad (6a)$$

where the last term is, of course, the change in Net Income. While we can now fairly closely approximate dz by $\Delta z \equiv z_t - z_{t-1}$, the choice of t or $t-1$ as the subscript for the non-differenced variables is arbitrary. By convention the prices would carry a $(t-1)$ subscript while (t) would be used for the quantity. Naturally, there is no compelling reason not to alter the convention.

Another method of deriving (6a) but this time with the time dimension of the variables exactly specified is to begin with the technology

$$F(\underset{\sim}{Q},\underset{\sim}{X},t) = 0 \qquad ; \qquad \underset{\sim}{z} \text{ is a vector}$$

where F is quadratic and by Diewert's Quadratic Lemma (Diewert (1976)) we get

$$\tfrac{1}{2} \sum(F_{Qt} + F_{Q,t-1})\Delta Q = \tfrac{1}{2} \sum(F_{Xt} + F_{X,t-1})\Delta X + \Delta TFP \qquad ;$$
$$\Delta TFP = TFP_t - TFP_{t-1} \quad . \qquad (7)$$

From profit maximization

$$F_Q = P \quad \text{and} \quad F_X = W$$

we can rewrite (7) as

$$\tfrac{1}{2} \sum(P_t - P_{t-1})\Delta Q = \tfrac{1}{2} \sum(w_t - w_{t-1})\Delta X + \Delta TFP \quad . \qquad (8)$$

$R \equiv C$ implies that $\Delta R \equiv \Delta C$, from the Quadratic Lemma

$$\Delta R = \tfrac{1}{2} \sum(P_t + P_{t-1})\Delta Q + \tfrac{1}{2} \sum(Q_t + Q_{t-1})\Delta P \tag{9}$$

$$\Delta C = \tfrac{1}{2} \sum(w_t + w_{t-1})\Delta X + \tfrac{1}{2} \sum(X_t + X_{t-1})\Delta w \tag{10}$$

and substituting (9), (10) and $\Delta R \equiv \Delta C$ into (7) we get

$$\Delta TFP = -\tfrac{1}{2} \sum(Q_t + Q_{t-1})\Delta P + \tfrac{1}{2} \sum(X_t + X_{t-1})\Delta w \cdot$$

Separating the inputs as per equation (6a) we can now write

$$\begin{aligned}\Delta TFP + \tfrac{1}{2}(Q_t + Q_{t-1})\Delta P + \tfrac{1}{2}(T_t + T_{t-1})\Delta K = {} & [\tfrac{1}{2}(L_t+L_{t-1})\Delta w + \tfrac{1}{2}(M_t+M_{t-1})\Delta m] \\ & + [\tfrac{1}{2}(K_t+K_{t-1})(\Delta r_{1t}+\Delta r_{2t}+\Delta r_{3t}) \\ & + \tfrac{1}{2}\{(r_{1t}+r_{1,t-1}) + (r_2+r_{2,t-1}) \\ & \quad + (r_{3t}+r_{3,t-1})\}\Delta K] \\ & + [(K_t+K_{t-1})\Delta r_{4t} + (r_{4t}+r_{4,t-1})\Delta K)\end{aligned} \tag{11}$$

The last expression, except for the form of the coefficients, which are now explicit, is identical to equation (6a). While (11) may be less arbitrary it is not entirely clear that it is superior for every choice of coefficient variable in (6a).

While the above model provides an extremely useful disaggregation of the financial/accounting income statement, it must be noted that nowhere in the model is anything said about the adequacy of the NI , upon which the relative impact of all the other items is being measured. Given that it is a residual in the cost of capital after payments to depreciation, debt service and taxes, we are led to believe that, within the context of the model, the return to invested capital, i.e., NI, is in fact also identically equal to its cost. Until now, the cost of capital has been defined residually, but this may not be useful in the long run, since it does not reflect the option the firm has to invest its internal generated fund in the capital market. Nevertheless, despite that drawback, this type of income statement presentation can only be a major improvement over the standard format since above all, it isolates the impact of inflation. In addition, while it presents the crucial information to be garnered from a knowledge of the relative impacts of TFP and individual price movements, it preserves all the key information normally found on an income statement including, of course, the critical net income results, now decomposed into inflationary price movements and productivity increases.

III. UNIPA (Unconstrained NIPA)

i) the model

The corner stone of the NIP model is $R = C$. However, once the cost of capital is defined exogenously, then it does not necessarily follow that R equals C. The cost of capital in the NIPA, through r_4, is whatever balances costs and revenues and nothing in the NIPA analysis prevents r_4 from being very high or very low or even negative, reflecting a very good or a very poor performance on the part of the firm. Evidently a good or a poor performance is a concept which has to be defined. This is not a problem since it has a common sense meaning which is formalized in economic analysis as the opportunity cost. To the extent the firm could dispose in some alternative way of its capital stock so as to receive at most a return of $\rho_t K_t$, then any return below ρ_t will be a poor performance since the firm could reorganize its resources to earn ρ_t . Similarly, a return above ρ_t will be a good performance. Now if we define the cost faced by the firm, where $\rho_t K_t$ is the opportunity cost of capital, such that

$$C_t' \equiv C(\rho_t) = w_t L_t + m_t M_t + \sum_{\ell=1}^{3} r_{\ell,t} K_t + \rho_t K_t \qquad (12)$$

then

$$PL \equiv R - C'$$

where PL is the profit or loss due to the unanticipated returns (positive or negative), and C' represents all incurred costs with the capital cost portion including the required return to invested capital. Nevertheless, since the definition of productivity still holds, given

$$\dot{TFP} = \dot{Q} - \dot{X}$$

then

$$\dot{TFP} = \dot{W} - \dot{P}$$

if, and only if $\dot{PL} = 0$. That is, PL is the repository of all deviations from plan. Noting that the plan was based on PL = 0 , i.e., $R^* = C^*$,

$$PL = (R-R^*) - (C'-C^*)$$

where the asterisks denote desired or planned values. Considering that $R = PQ$ and $C' = W'X'$ the complete revised NIPA expression can now be derived from $R \equiv C' + \dot{PL}$, which can be rewritten in terms of proportional charges,

$$\dot{R} = \frac{C'}{R}\dot{C}' + \frac{PL}{R}\dot{PL} \quad .$$

From $\dot{R} = \dot{Q} + \dot{P}$; $\dot{C}' = \dot{X}' + \dot{W}'$ and the expression for $\dot{PL}$ above, as well as the fact that Q, P, X' and W' are indices of output quantities, output prices, input quantities and input prices, respectively, it follows that

$$\dot{Q} + \dot{P} = \frac{C'}{R}[\dot{X}' + \sigma_w\dot{w}' + \sigma_m\dot{m}' + \sigma_r\dot{r}'] + \frac{PL}{R}\dot{PL}$$

$$\left[\dot{Q} - \frac{C'}{R}\dot{X}'\right] + \dot{P} + \frac{C}{R}\sigma_r\dot{K}' = \frac{C'}{R}[\sigma_w\dot{w}' + \sigma_m\dot{m}'] + \frac{C'}{R}[\sigma_r\dot{r}' + \sigma_r\dot{K}] + \frac{PL}{R}\dot{PL} \quad (13)$$

where $\dot{PL} = 1$ and where we recognize $[\sigma_r\dot{r} + \sigma_r\dot{K}']$, with one difference, as the combination of depreciation, tax and financial and Net Income growths of the standard NIPA analysis. The difference is that the weights σ_i are based on C' which is equal to R if and only if $PL = 0$.

Finally in order to make (13) operational it must be transformed. We expand (13) to

$$\left[\frac{PQ}{R}\frac{dQ}{Q} - \frac{W'X'}{R}\frac{dX'}{X'}\right] + \frac{PQ}{R}\frac{dP}{P} + \frac{C'}{R}\;\frac{r'K'}{C'}\frac{dK'}{K'} = \frac{C'}{R}\left[\frac{w'L'}{C'}\frac{dw'}{w'} + \frac{m'M'}{C'}\frac{dm'}{m'}\right]$$

$$+ \frac{C'}{R}\frac{r'K'}{C'}\left[\frac{dr'}{r'} + \frac{dK'}{K'}\right]$$

$$+ \frac{PL}{R}\frac{dPL}{PL} \qquad (14)$$

Multiplying through by R and cancelling all other denominator terms and replacing the continuous differential sign 'd' by 'Δ' , we get,

$$[P\Delta Q - W'\Delta X'] + Q\Delta P + r'\Delta K' = [L'\Delta w' + M'\Delta m'] + [K'\Delta r' + r'\Delta K'] + \Delta PL \quad (15)$$

Expression (15) is now amenable to tabulation in dollar terms for management. The only remaining question, as with the NIPA analysis above, pertains to the choice of (t) or (t-1) as the subscript for the coefficient variables. We could of course have derived the same expression using Diewert's Quadratic Lemma, except that then the coefficient variables would have been exactly defined to give $\frac{1}{2}(Z_t + Z_{t-1})\Delta Y$.

ii) <u>post-mortem utilisation of UNIPA</u>

The UNIPA model is here modified to do a post-mortem analysis in which we recognize that deviations from plan are an unavoidable phenomena which will generate positive or negative unanticipated earnings (UE) . Whereas ex ante the firm will plan to earn a "desired" return, ex post realities will usually differ from anticipations. It should be noted that when we refer to "desired" returns we mean those amounts required to exactly offset all costs, including labour, capital and materials. As before, the firm plans for revenues which, after paying labour, intermediate goods and services suppliers, depreciation expenses, financial obligations and taxes, will leave a residual to "adequately" compensate the providers of equity capital. However, as is the nature with any residual, in situations of uncertainty, it will equal its planned level, in the short run, only by coincidence. In this version of the NIPA model we both account for as well as explain these deviations from plan. The accounting identity $R \equiv C + PL$ is now replaced by

$$UE = (R - R^*) - (\tilde{C} - C^*)$$

However the exogenous return on capital is now defined not in terms of the opportunity cost the firm would reach were it to shift its operation but rather in terms of the rate of return it was expected to reach when it developed its plan. This rate will be denoted by γ_t, such that

$$\tilde{C}_t = \tilde{C}(\gamma_t) = w_t L_t + m_t M_t + \left(\sum_{\ell=1}^{3} r_{\ell,t} + \gamma_t \right) K_t \quad .$$

For simplicity, let $\tilde{C}_t = W_t X_t$ where W_t and X_t are appropriate price and quantity input indexes, then

$$dUE = R\dot{Q} + R\dot{P} - R^*\dot{Q}^* - R^*\dot{P}^* - \tilde{C}\dot{X} - \tilde{C}\dot{W} + C^*\dot{X}^* + C^*\dot{W}^* \quad .$$

Dividing through by R, we obtain the unanticipated earnings as a ratio expressed in terms of the realized revenue:

$$\frac{dUE}{R} = \dot{Q} + \dot{P} - \left(\frac{R^*}{R} \right)\dot{Q}^* - \left(\frac{R^*}{R} \right) \dot{P}^* - \left(\frac{\tilde{C}}{R} \right) \dot{X} - \left(\frac{\tilde{C}}{R} \right) \dot{W}$$

$$+ \left(\frac{R^*}{R} \right) X^* + \left(\frac{R^*}{R} \right) \dot{W}^*$$

where we used $R^* = C^*$ through which γ_t was defined.

Denoting the inverse of the realized revenue as a ratio of the planned revenue by γ , i.e., $\gamma = R^*/R$, and regrouping terms to isolate the TFP components, we obtain, noting that $UE^* = 0$,

$$\frac{UE}{R} = [\dot{TFP} - \gamma\ \dot{TFP}^*] + [\dot{P} - \gamma\ \dot{P}^*] - [\dot{W} - \gamma\ \dot{W}^*] + \frac{UE}{R}\dot{\tilde{C}}$$

where we have used $\dot{\tilde{C}} = \dot{X} + \dot{W}$.

Finally

$$\frac{UE}{R} = (1 - \dot{\tilde{C}})^{-1}\{[\dot{TFP} - \gamma\ \dot{TFP}^*] + [\dot{P} - \gamma\ \dot{P}^*] - [\dot{W} - \gamma\ \dot{W}^*]\}$$

i.e., the unanticipated earning as a ratio of revenue is a weighted sum of the difference between the planned and the realized values. The first term in brackets is that proportion of the unanticipated earnings due to the difference between planned and actual productivity growth while the second and third terms reflect the degrees to which planned and actual price recovery differs. It is to be noted that the planned rates of growth are corrected for the error in revenue forecast, γ . The entire expression of course reflects the degree to which the productivity divergence and price recovery divergence offset each other. These can of course be broken down into all the same elements as the actual UNIPA statement.

The post-mortem analysis adds a new dimension to analysis of the net income in that it enables one to study the impact of the various forecasting errors, be they of exogenous variables such as w_t, m_t, ... or of endogenous terms such as L_t, P_t, ... through costs and revenues on the income statement. For instance the impact of a strike which might significantly lower L_t but which may be associated with an unforeseen wage settlement which, in turn, might increase significantly w_t can now be traced, ...

By decomposing as in the NIPA and UNIPA analysis $[\dot{W} - \gamma_t\dot{W}^*]$, we obtain

$$[\dot{W}_t - \gamma_t\dot{W}_t^*] = [\sigma_{L,t}\dot{w}_t - \gamma_t\sigma_{L,t}^*\dot{w}_t^*] + [\sigma_{m,t}\dot{m}_t - \gamma_t\sigma_{m,t}^*\dot{m}_t^*]$$

$$+ \sum_{\ell=1}^{3}[\sigma_{r,t}\varepsilon_{\ell,t}\dot{r}_{\ell,t} - \gamma_t\sigma_{r,t}^*\varepsilon_{\ell,t}^*\dot{r}_{\ell,t}^*]$$

$$+ [\sigma_{r,t}\varepsilon_{4,t}(d\ln \gamma_t K_t) - \gamma_t\sigma_{r,t}^*\varepsilon_{4,t}^*(d \ln \gamma_t K_t^*)]$$

$$- [\sigma_{r,t}\varepsilon_{4,t}\dot{K}_t - \gamma_t\sigma_{r,t}^*\varepsilon_{4,t}^*\dot{K}_t^*]$$

and substituting in the previous equation, we have

$$(\frac{UE}{R}) = (1 - \dot{\tilde{C}})^{-1}\{([\dot{TFP} - \gamma\ \dot{TFP}^*] + [\dot{P} - \gamma\dot{P}^*] + [\sigma_{r,t}\varepsilon_{4,t}\dot{K}_t - \gamma_t\sigma_{r,t}^*\varepsilon_{4,t}^*\dot{K}_t^*]$$

$$- ([\sigma_{L,t}\dot{w}_t - \gamma_t\sigma_{L,t}\dot{w}_t^*] + [\sigma_{m,t}\dot{m}_t - \gamma_t\sigma_{m,t}\dot{m}_t^*]$$

$$+ \sum_{\ell=1}^{3}[\sigma_{r,t}\varepsilon_{\ell,t}\dot{r}_{\ell,t} - \gamma_t\sigma_{r,t}^*\varepsilon_{\ell,t}^*\dot{r}_{\ell,t}^*])$$

$$- ([\sigma_{r,t}\varepsilon_{4,t}(d\ln \gamma_t K_r) - \gamma_t\sigma_{r,t}^*\varepsilon_{4,t}^*(d\ln \gamma_t K_t^*)]))\}$$

The three terms on the RHS are respectively the positive NIPA factors of productivity, output price and capital growth, the negative NIPA factors of errors in forecasting in wages, price of materials, depreciation, taxes and financial charges, and finally the weighted impact on net income of an error in the construction program.

In expanding the elements of (3) as we did for the standard UNIPA analysis, each individual item from the NIPA statement can be matched with its own unique variance. In essence we would have something resembling:

Plan	Actual	Variance
Positive Factors		
TFP	TFP	Due to TFP
+ Output Price Changes	+ Output Price Changes	Due to Output Price Changes
+ Capital Growth	+ Capital Growth	Due to Capital Growth
Negative Factors		
- Input Price Changes	- Input Price Changes	Due to Input Price Changes
- Capital Cost Changes (excluding NI)	- Capital Cost Changes	Due to Capital Cost Changes
= NI	= NI	UE
UE = 0	U = NI plan - NI actual ≠ 0	

IV. Integrated Planning Model

a) Introduction

The two versions of NIPA, presented above, while providing a good analytical framework for the intelligent evaluation of bugetary plans, are essentially ex post models. NIPA intervenes in the budgetary process in a sequential manner, taking an active role only after the laborious planning exercise produces its game plan. At that juncture NIPA analyses the budget's implicit productivity performance, which may or may not justify another round of the planning process. Given the scope of the bugetary process in any large firm, it is unlikely that a bad productivity picture, along with good built-in financial results, will move the planners to modify an already overly complex structure. The most natural solution to this dilemma would be to ensure that NIPA results are always favourable. This can be done by including productivity as an explicit consideration during the planning process. Such a model is the subject of this section. We will present a model which can be used to develop a complete, theoretical, corporate plan (budgetary and otherwise), explicitly incorporating all essential physical and financial targets such as return to investment and productivity. In this way, top management, who ultimately have to approve any budget, will have available a set of guidelines, incorporating all essential corporate objectives, through which to more closely guide the development of the actual budgetary process. They will be in a position to set spending guidelines that, if exceeded, will ensure that some or all of the target constraints are not fulfilled.

It is a mixed model, using econometrics only when the constraints of a pure accounting approach detract significantly from its ability to mirror the real world. In particular, as well be seen below, econometrics are used to estimate the relative input factor cost shares which ultimately translate into the basic technological ratios of the production process.

The major advantage of the following model lies in its simultaneous approach to the planning problem. In most purely financial planning models the distinct identifiable input sector is, to a large extent, independently sized and then fitted into the framework of certain corporate constraints, which include the financial rate of return. It is of course only by coincidence that such a process will end with a perfect fit after a first attempt. Some of the items will be recycled and returned for a new round of integration. We do not mean to imply that there is no prior interaction between the various sectors or that productivity is not an important consideration, only that the interactions and productivity considerations are partial in nature.

If we look at Figure 1, which assumes a capital intensive firm, thus placing a large importance on the capital budgetary process, we can trace the evolution (in very general terms) of a corporate budgetary plan. The most important driving forces are prior and present period demand forecasts. The former creates a requirement for ongoing capital projects, pretty well divorced from present demand conditions, while the latter determines present and longer term capital projects as well as, to a certain extent, replacement requirements. "Other" reasons for increasing the capital budget vary from industry to industry. In telecommunications, for example, international standards and interface exigencies would play significant roles. Regulated industries, in general, would find their capital budgets subject to pressures other than market demand. Ultimately, all the capital requirements are evaluated at current asset prices and a capital budget is derived.

The technological characteristics of the capital budget create part of the demand for the other input factor. These include the general categories of labour and other expenses (henceforth to be referred to as "materials"). They comprise such items as maintenance, direct operating labour, rental of

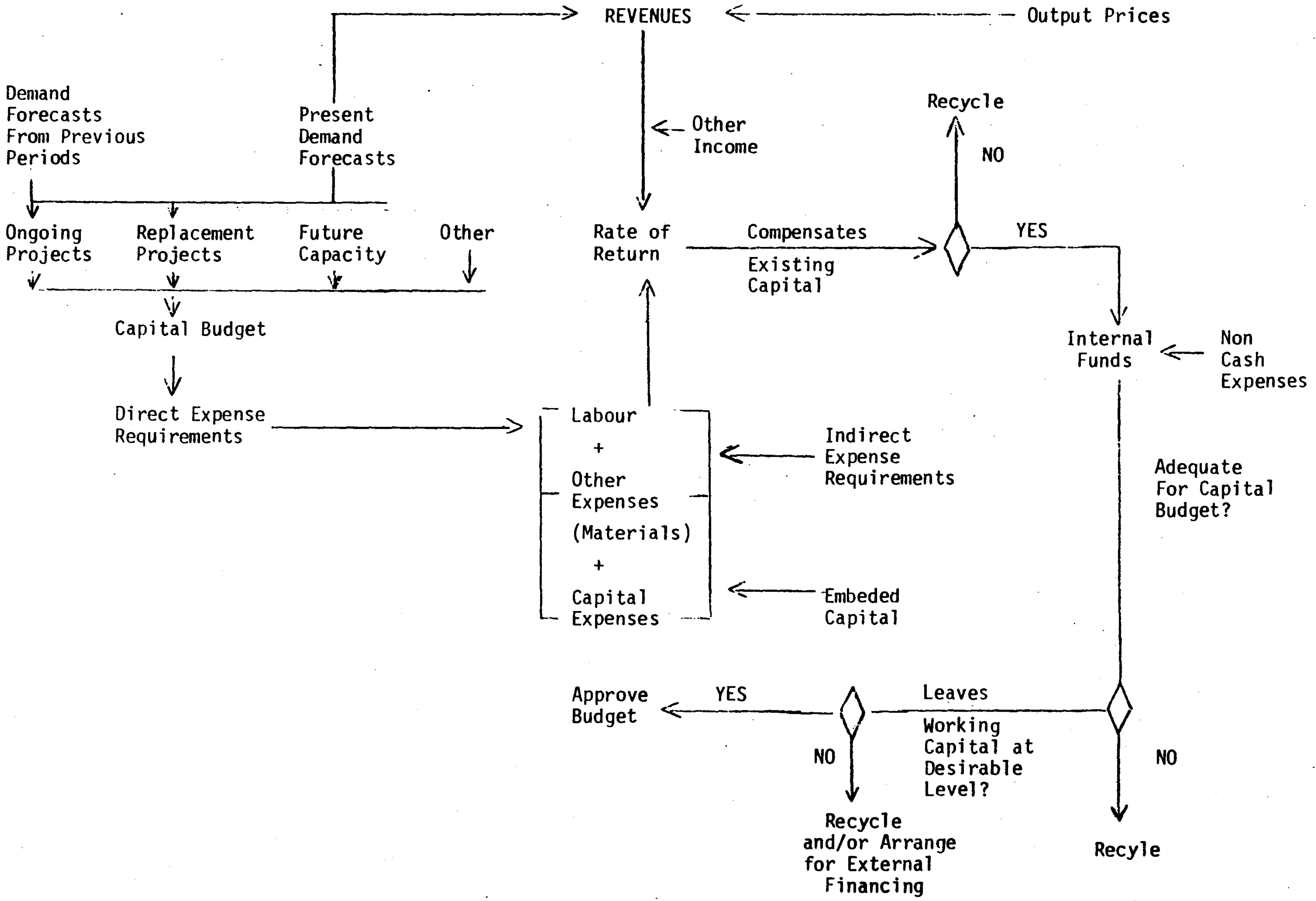

REVENUES
Output Prices
Demand Forecasts From Previous Periods
Present Demand Forecasts
Other Income
Recycle
NO
Ongoing Projects
Replacement Projects
Future Capacity
Other
Rate of Return
Compensates Existing Capital
YES
Capital Budget
Internal Funds
Non Cash Expenses
Direct Expense Requirements
Labour
+
Other Expenses (Materials)
+
Capital Expenses
Indirect Expense Requirements
Embeded Capital
Adequate For Capital Budget?
Approve Budget
YES
Leaves Working Capital at Desirable Level?
NO
NO
Recycle and/or Arrange for External Financing
Recyle

Figure 1

facilities, etc. In addition, the various components of the capital budget, as well as embedded capital, determine the value of capital costs. These include depreciation expenses, interest payments, taxes and, ultimately, the value of earnings applicable for dividend payments to equity holders. This is the residual, after payment to all factors, including debt capital, that ultimately compensates the owners of the firm. When calculated as a percentage of total invested catpial, then it is known as the rate of return.

It is within this capital/other factor interaction that "quasi" partial productivity considerations make their first appearance. Quasi, because these are really measures of worker efficiency rather than true overall productivity measurements. They are industrial engineering measures such as "work units" which compare performance against established standards. They take no account of the negative contribution to overall productivity when capital is used to increase work units per unit of time. Naturally, the link between these measures and overall corporate performance is difficult to establish.

The other determinants of total expenses are only indirectly related to capital budgeting and are determined more as a result of overall business size and prosperity. These include all those luxury factors such as marketing, training, special studies, etc. That is, the entire set of indirect, non-operating expenses.

Total revenues, including forecast demand at given prices and other, non-operating income, are combined with the total value of current input to determine the residual and, ultimately, the rate of return. If the RIR is inadequate, in that it either fails to compensate existing capital at a fair rate or does not cover all capital expenditures without excessive external financing requirements then there occurs a budgetary recycling process where all or part of the plan is altered. Usually it is the latter, concentrating on the expense rather than capital budget items. Corrective action may include labour cuts, material cuts, output price changes and, as a last resort, capital budget cuts.

Significant by their absence are the aspects of simultaneity and some overall explicit recognition of productivity. The advantage of simultaneously calculating all the unknowns are obvious, but what are the advantages of including productivity? Simply that the implied technological relationship of a production function, as embodied in the explicitly reocgnized productivity number allows for a combination of inputs, given the output, that is in some sense optimum. This optimum provides an additional constraint to the general planning problem which serves to narrow the choice between the various input options to more manageable proportions.

b) The Model

The model postulates the existence of some cost function

$$C = g(w,m,r,Q,t) \tag{1}$$

where w = the price of labour

m = the price of materials (or intermediate expense items)

r = the periodic (say, annual) cost of using the capital stock. It includes:

δ = depreciation rate

ϕ = the rate of taxation

θ = the return to outstanding debt

π = the return to equity

Q = the volume of output produced

t = the technology indicator.

From (Denny, Fuss & Everson (1979)) and (Denny, de Fontenay & Werner (1980)) we totally differentiate the cost function with respect to time to yield:

$$\frac{dC}{dt} = \frac{\partial g}{\partial w}\frac{\partial w}{\partial t} + \frac{\partial g}{\partial m}\frac{\partial m}{\partial t} + \frac{\partial g}{\partial r}\frac{\partial r}{\partial t} + \frac{\partial g}{\partial Q}\frac{\partial Q}{\partial t} + \frac{\partial g}{\partial t} \qquad (2)$$

Rearranging through division by C and from Sheppard's Lemma setting $\frac{\partial g}{\partial q_i} = X_i$; $q_i = w,m,t$ and $X_i = L, M$ and K respectively, we get

$$\frac{1}{C}\frac{dC}{dt} = \sigma_w \frac{dw}{dt}\frac{1}{w} + \sigma_m \frac{dm}{dt}\frac{1}{m} + \sigma_r \frac{dr}{dt}\frac{1}{r} + \frac{\partial g}{\partial Q}\frac{Q}{C}\left(\frac{\partial Q}{\partial t}\frac{1}{Q}\right) + \frac{1}{C}\frac{\partial g}{\partial t} \qquad (3)$$

where $\sigma_i = \frac{q_i X_i}{C}$; for $q_i = w, m, r$ and $X_i = L,M,K$.

which are the cost shares of each input and

L = manhours of input

M = materials inputs

K = the stock of physical capital .

From the definition of costs

$$C = wL + mM + \bar{r}K \quad .$$

By totally differentiating with respect to time and rearranging we get

$$\sum_{i=1}^{3} \frac{q_i X_i}{C}\frac{dq_i}{dt}\frac{1}{q_i} = \frac{1}{C}\frac{dC}{dt} - \sum_{i=1}^{3} \frac{q_i X_i}{C}\frac{dX_i}{dt}\frac{1}{X_i}$$

or

$$\sum \sigma_i \frac{dq_i}{dt_i}\frac{1}{q_i} = \frac{1}{C}\frac{dC}{dt} - \sum \sigma_i \frac{dX_i}{dt}\frac{1}{X_i} \quad .$$

Substituting this into (3) above we get

$$-\frac{1}{C}\frac{\partial g}{\partial t} = \left(\frac{\partial g}{\partial Q}\frac{Q}{C}\right)\left(\frac{\partial Q}{\partial t}\frac{1}{Q}\right) - \sum \sigma_i\left(\frac{dX_i}{dt}\frac{1}{X_i}\right) \quad .$$

If we assume that the cost elasticity, $\frac{\partial g}{\partial Q}\frac{Q}{C}$, is approximately equal to 1 over the period under consideration, then

$$-\frac{1}{C}\frac{\partial g}{\partial t} = \frac{\partial Q}{\partial t}\frac{1}{Q} - \sum \sigma_i\left(\frac{dX_i}{dt}\frac{1}{X_i}\right)$$

where the right hand side is the shift in the production function due to technology, and, by definition, is equal to the change in total factor productivity, $\dot{TFP}$ and

$$\dot{TFP} = \frac{\partial Q}{\partial t}\frac{1}{Q} - \sum \sigma_i\left(\frac{dX_i}{dt}\frac{1}{X_i}\right) \quad . \qquad (4)$$

We may rewrite (4) in discrete form:

$$\dot{TFP} = (\ln Q_1 - \ln Q_0) - \sum \tfrac{1}{2}(\sigma_{i1} + \sigma_{i0})(\ln X_{i1} - \ln X_{i0}) \qquad (5)$$

where $\sigma_i = \frac{1}{2}(\sigma_{i1} + \sigma_{i0})$. We can now rearrange equation (5) so that it can be solved for any one of the X_i , say K , then:

$$\ln K = \left[\ln\left(\frac{Q_1}{Q_0}\right) + \sigma_L \ln L_0 + \sigma_M \ln M_0 + \sigma_K \ln K_0\right]$$

$$+ (1-\sigma_K)\left[\ln\left(\frac{K_1}{L_1}\right)\right] - \sigma_M\left[\ln\left(\frac{M_1}{L_1}\right)\right] - \dot{TFP} \qquad (6)$$

Equation (6) has several unknowns and is at present not soluble. From the cost function g as a translog we can derive equations for each of the cost shares σ_{i1} .[5]

$$\sigma_{L1} = \alpha_L + \alpha_{LL} \ln w_1 + \alpha_{LM} \ln m_1 + \alpha_{LK} \ln r_1 + \alpha_{LQ} \ln Q_1 + \alpha_{Lt} t$$

$$\sigma_{M1} = \alpha_M + \alpha_{ML} \ln w_1 + \alpha_{MM} \ln m_1 + \alpha_{MK} \ln r_1 + \alpha_{MQ} \ln Q_1 + \alpha_{Mt} t$$

$$\sigma_{K1} = \alpha_K + \alpha_{KL} \ln w_1 + \alpha_{KM} \ln M_1 + \alpha_{KK} \ln r_1 + \alpha_{KQ} \ln Q + \alpha_{Kt} t$$

In the above system since $\sum \sigma_{i1} = 1$, we need only estimate any two and then solve for the third set of coefficients from the following conditions

$$\sum_i \alpha_i = 1 \quad ; \quad \sum_i \alpha_{ij} = 0 \quad ; \quad \sum_i \alpha_{iQ} = 0 \quad ; \quad \sum_i \alpha_{it} = 0$$

For our model we assume that w_1, m_1 and t are known and r is unknown. Therefore, in order to get estimates for the α_i and α_{ij} , we estimate the equation only to period 0 . Then the $\sigma_{i1} = h(r)$.

Further, from the definition:

$$\sigma_{i1} = \frac{q_{i1} X_i}{C}$$

we can find the ratios:

$$\frac{K_1}{L_1} = \frac{w_1}{r_1} \frac{\sigma_{K1}}{\sigma_{L1}} \qquad \text{and} \qquad \frac{M_1}{L_1} = \frac{w_1}{m_1} \frac{\sigma_{M1}}{\sigma_{L1}} \qquad (7)$$

where the ratios are each functions, by virtue of the share equations, only of r . We now have two unknowns, r and K and one equation, (6).

Given that our aim is to integrate our model directly into the corporate planning routine, the cost of capital r, which has economic meaning must be related to the financial cost of capital, r^* where

$$r^* = \delta + \lambda\theta + (1-\lambda)(1-\phi)\pi \qquad (8)$$

where λ is the proportion of total financial capital in the form of debt. The relation then can be postulated as:

$$rK = r^*K^B \qquad (9)$$

where K^B = the net original value of physical capital which, by definition equals the value of financial capital. In addition we also have, by definition:

$$A_0 = K_1^B - K_0^B + R_1(R_1-R_1^*) = K_1^B - (K_0-K_1^*)$$

$$A_1 = q_1(K_1-K_0) + R_1^*$$

where A_1 = the value of gross additions to the plant

R_1^* = the value of retirements that are actually replaced

R_1 = the value of retirements .

We can now derive the following relation:

$$(K_1-K_0) = (r_1-r_1^*q_1)^{-1}\{-r_1K_0 + r_1^*[K_0^B - (R_1-R_1^*)]\}$$

Of course, if all retired plants are ultimately replaced, either by exact reproductions or new technology then $(R_1-R_1^*) \approx 0$ and

$$(K_1-K_0) = (r_1-r_1^*q_1)^{-1}[-r_1K_0 + r_1^*K_0^B] \qquad (10)$$

Equations (6) and (10) now form a system of two equations in the two unknowns r_1 and K_1. All the other unknowns of the general planning problem can now be derived from the solution to the system (6) and (8). Given a value for r_1, the share variable σ_{i1} assume values which, from (7), produce solutions for L_1 and M_1. This, along with the prices w_1, m_1 and r_1, puts a value on total cost which of course implies a total revenue requirement. Thus, we can see, that given the key constraints of demand forecasts, rate of return requirements and desired productivity growth we have calculated a cost equation whose components all embody the constraints:

$$C = r_1K_1 + w_1L_1 + m_1M_1$$

Further, taking account of the accounting identity whereby total revenues should be identically equal to total costs,

$$R \equiv C$$

$$PQ \equiv C$$

then we have a required price level for output as well. For all the other details of a full-blown financial plan we can use equations (8) and (9) to calculate depreciation expenses, taxes, interest payments, the various balance sheet items, source and uses statements and so on.

V. Conclusion

The notion that productivity is an important part of business success, as stated at the outset, may not be a new concept, but to incorporate it explictly into an overall corporate/budgetary plan is. In this paper we have demonstrated two ways of going about this integration. The first, involve more of a static budgetary analysis in the form of NIPA and UNIPA. They take, as given, the financial/accounting information in any plan, and compute the relative impact of productivity, among other variables, on the growth in Net Income, which, after all, is the firm's ultimate measure of management success. While NIPA imposes the constraint that all returns to factor are always identically equal to their costs, UNIPA does not.

The other method of introducing productivity into the corporate/budgetary planning exercise involves a direct intervention in the process. TFP itself becomes a target variable and thus a parameter in the actual derivation of a complete guideline plan. Based on the desired levels of productivity, financial return and production (to meet anticipated demand), the planning model simultaneously calculates all the relevant variables of an entire plan which includes the income statement, balance sheet and funds flow information. While it does provide all the pertinent operating information the results of the model are not meant to replace the normal bottom-up planning process. Instead they offer a complete set of guidelines for upper management on the values of key operating indicators such as employee expenses, manhours, capital budgeting, etc. which, if not attained, will imply the untenability of management's key task targets, including financial return to investment, production level and productivity gains.

Footnotes

1. The original work on the management use of TFP by a firm must be credited to the Electricité de France (EDF), and its surplus analysis (Reimeringer, 1980) is the forerunner of all NIPA models. Certain multinational corporations, such as IBM, Xerox, ... are known to use TFP measures as general guidelines and DRI is in the process of formalizing such an idea. In 1977, Teleglobe Canada and the British Columbia Telephone Company organized two symposia at which a number of Canadian telecommunicationscarriers came together to discuss the concept and measurement of TFP. Nevertheless, the active and systematic use of TFP as a management tool, introduced analytically in the management process, but for EDF, appears to have been pioneered by telecommunications carriers, with Teleglobe Canada and AT&T in the process of incorporating it in the formal budgeting and planning process and with Bell Canada developing similar internal uses. In addition, two other Canadian telecommunications carriers have on-going productivity studies, British Columbia Telephone Company and Alberta Government Telephone. Finally, nine Canadian telecommunications carriers are participating with the Canadian Department of Communications in a major productivity project, which has, as one of its goals the development of management uses of TFP analysis.

REGULATION OF TELECOMMUNICATION RATES

USING A SIMPLE FORMULA PROCEDURE

BY

RAY J. GOODIER

PART A: - DEVELOPMENT OF "THE RAFP SIMPLE FORMULA"

A-I. INTRODUCTION

To make the regulation of telecommunications carriers more responsive to inflationary pressures resulting in the need for more frequent, lengthy, and expensive public rate hearings, the Canadian Transport Commission (CTC), on 15 August 1974, proposed a Rate Adjustment Formula Procedure (RAFP), and requested interested parties to submit their comments.[1]

Two years later on 7 September 1976, the Canadian Radio-television and Telecom. Commission (CRTC), having taken over the Telecom. authority of the CTC, and having considered the arguments raised for and against RAFP, announced their decision to "suspend the rate adjustment formula proceedings"[2] due to a concern over "a number of technical difficulties",[3] but primarily because the CRTC considered that the "carriers under its jurisdiction must continue to be accountable through public hearings for all general rate increases".[4]

Notwithstanding the CRTC's justification for suspending RAFP, I believe RAFP should be reconsidered since carriers are now facing the same high inflation they experienced in 1974 when RAFP was first proposed. It is significant that such high inflation did not exist at the time RAFP was suspended due to the fact that the Anti-Inflation Program had been in effect for one year and was expected to restrict inflation for many years to come. Indeed, the CRTC specifically stated in its announcement suspending RAFP that "the regulatory environment was significantly changed for both the carriers and the Commission with the introduction of the Anti-Inflation Program in October, 1975".[5] Clearly, in view of the failure of the Anti-Inflation Program, and the current high rate of inflation, RAFP deserves to be reappraised.

In the hope that others agree, I will present in this paper my analysis of RAFP, and will show the development of a remarkable "RAFP Simple Formula" (not previously considered) that satisfies all the assessment criteria laid down by the CTC, yet requires neither the estimation of a rate of inflation, nor the calculation of a rate of productivity gain.

A-II. ANALYSIS OF RAFP

The essence of RAFP is the application of an RAFP formula to a recent Test Period (e.g., last complete fiscal year) and preceding Base Period to arrive at a carrier's required Revenue Adjustment, i.e., the annual dollar value of the future rate adjustment permitted by the regulatory body.

Under the CTC proposal, the RAFP formula must meet three criteria as found in the attachment to CTC Order No. T-474 of 15 August 1974.

1. The formula selected should compensate the carriers for the uncontrollable changes in costs.[6]

2. Productivity gains can be used by the carrier to offset some of the uncontrollable costs it incurs.[7]

3. The cost increases associated with the growth component are expected to be recovered through increased revenues and there will be no rate adjustment for the growth component.[8]

Significantly, these criteria ralate to the three factors: INFLATION, PRODUCTIVITY GAIN, and GROWTH which together are sufficient to account for all changes in cost experienced by the carriers. Specifically, it may be shown that total cost will change in response to the three factors according to the following relationship:

$$\begin{pmatrix}\text{The Test}\\ \text{Period Cost}\end{pmatrix} = \begin{pmatrix}\text{The Base}\\ \text{Period Cost}\end{pmatrix} \times \frac{\text{(Growth Index) (Inflation Index)}}{\text{(Productivity Gain Index)}}$$

or,

(A1) $TC = BC(1+g)(1+c)/(1+p)$

Where: TC = test period cost,

BC = base period cost,

$1+g$ = growth index,

$1+c$ = inflation index, and

$1+p$ = productivity gain index.

With respect to terms c, g, and p, some further explanation is required. Term c may be defined as the overall rate of inflation experienced by the carrier, or alternatively as "the rate of uncontrollable increase in costs."[9]

Term g may be defined as the overall rate of output growth, where output is taken to mean revenues adjusted to a common tariff (e.g., base period and test period revenues both expressed in terms of base period tariff). Thus, we may state:

(A2) $1+g = TR/BR$

Where: $1+g$ = growth index,

TR = test period revenues adjusted to base period tariff, and

BR = base period revenues.

Finally, term p may be defined as rate of productivity gain, or rate of change in Productivity Index (PI), where Productivity Index is defined as Total Constant Dollar Revenue divided by Total Constant Dollar Cost, i.e., (A3) $PI = Output/Input$

Where: PI = productivity index,

Output = revenues expressed in terms of base period tariff,

Input = costs expressed in terms of base period dollars.

For the base period and test period the above relationship becomes:

(A4) $BI = BR/BC$, and

(A5) $TI = \dfrac{TR}{TC/(1+c)}$

Where: BI = base period productivity index,

BR = base period revenues,

BC = base period costs,

TI = test period productivity index,

TR = test period revenues adjusted to base period tariff,

TC = test period costs,

1+c = inflation index, and

TC/(1+c) = deflated test period costs.

Since p is the rate of change in productivity index, then we may state:

(A6) $1+p = TI/BI$.

From equations (A2), (A4), (A5) and (A6) we can now derive equation (A1):

(A6) $1+p = TI/BI = \dfrac{\dfrac{TR}{TC/(1+c)}}{BR/BC} = \dfrac{BC}{TC}(TR/BR)(1+c)$, or

(A7) $1+p = \dfrac{BC}{TC}(1+g)(1+c)$

which, by further rearranging, becomes equation (A1):

(A1) $TC = BC(1+g)(1+c)/(1+p)$ as required.

Returning to the RAFP proposal, we may calculate the uncontrollable changes in costs, or Inflation Cost (IC), as the actual test period cost less the hypothetical test period cost in the absence of inflation:

$$\text{(A8)} \quad IC = TC - TC/(1+c)$$

Where: IC = inflation cost,

TC = test period cost,

$1+c$ = inflation index, and

$TC/(1+c)$ = test period cost in absense of inflation (i.e., deflated test period cost).

Similarly, we may calculate the dollar value of the rate adjustment, or Revenue Adjustment (RA), as actual test period cost less the hypothetical test period cost in the absence of inflation and productivity gain:

$$\text{(A9)} \quad RA = TC - TC(1+p)/(1+c)$$

Where: RA = revenue adjustment,

TC = test period cost,

$1+c$ = inflation index,

$1+p$ = productivity gain index, and

$TC(1+p)/(1+c)$ = test period cost in absence of inflation and productivity gain.

Also, we may derive a comparable equation for the savings due to productivity improvement, or Productivity Gain (PG), by first restating the CTC proposal as follows:

$$(\text{Revenue Adjustment}) = (\text{Inflation Cost}) - (\text{Productivity Gain}),$$

or:

(A10) $RA = IC - PG$ which may be rearranged to give

(A11) $PG = IC \qquad - RA$

$$= TC - \frac{TC}{(1+c)} - TC + TC\frac{(1+p)}{(1+c)} \quad \text{(from (A8) \& (A9))}.$$

Thus, (A12) $PG = \frac{pTC}{(1+c)}$

In simple terms, this equation indicates:

> Productivity Gain equals the rate of productivity gain times the deflated test period cost.

Now with respect to growth, CTC criterion No. 3 requires that there be no rate adjustment for the growth component. Thus, it is of interest to derive an equation for the growth component, or Growth Cost (GC).

Clearly, in the absence of inflation and productivity gain, i.e., a simple scale change, one would expect costs to rise in direct proportion to the output growth. Thus, one may state the growth cost formula directly:

(A13) $GC = gBC$

Where: GC = growth cost,

g = rate of growth, and

BC = base period cost.

(Note that the employment of spare capacity must be considered as a productivity gain since input remains constant while output increases. Conversely, the absence of productivity gain implies a fixed proportion of spare capacity, i.e., a simple scale change.)

To complete the analysis of the RAFP proposal, consider again that cost changes are entirely due to growth, productivity gain and inflation.

This consideration implies:

(A14) $TC = BC + GC + IC - PG$

Where: TC = test period cost,

BC = base period cost,

GC = growth cost,

IC = inflation cost, and

PG = productivity gain.

As a test we can make substitutions and derive equation (A1):

$$\text{(A14)} \quad TC = BC + GC + IC - PG$$

$$= BC + gBC + TC - \frac{TC}{(1+c)} - \frac{pTC}{(1+c)}, \text{ or}$$

$$\text{(A1)} \quad TC = BC(1+g)(1+c)/(1+p) \qquad \text{as required.}$$

A-III. RAFP SIMPLE FORMULA

Having developed the foregoing equations, we now come to the main purpose of this paper: the development of a simple formula for calculating revenue adjustment without reference to inflation or productivity gain.

Consider again equations (A1) and (A9):

(A1) $TC = BC(1+g)(1+c)/(1+p)$ and

(A9) $RA = TC - TC(1+p)/(1+c)$,

and note that equation (A1) may be rearranged:

(A15) $TC(1+p)/(1+c) = BC(1+g)$.

Finally, combining equations (A9) and (A15), we arrive at the following "RAFP Simple Formula":

$$\text{(A16)} \quad RA = TC - BC(1+g)$$

Where: RA = revenue adjustment,

TC = test period cost,

BC = base period cost, and

$1+g$ = output growth index.

Since the RAFP Simple Formula excludes terms c and p, this means that the revenue adjustment may be calculated without the need to determine either a rate of inflation, or a rate or productivity gain. This is important since the determination of these rates is undoubtedly the major source of those "technical difficulties"[10] that concern the CRTC.

One may wonder how inflation and productivity gain can be treated as irrelevant when each obviously has a distinct impact on costs. The answer lies in the fact that irrespective of their individual values their <u>combined effect</u> must have been unique to produce the test period cost actually experienced. More specifically, for given values of BC, TC, and g, the expression (1+p)/(1+c) must be a constant for all estimations of c.

The above follows from equation (A1):

$$\text{(A1)} \quad TC = BC(1+g)(1+c)/(1+p)$$

which may be transformed into

$$\text{(A17)} \quad \frac{1+p}{1+c} = \frac{BC}{TC}(1+g) = \text{a constant (given BC, TC and g).}$$

Since (1+p)/(1+c) is a constant, then p will appear as high or low depending on whether the estimate for c is correspondingly high or low.

It also follows, given values for TC, BC, and g, that equation (A9), i.e.,

$$\text{(A9)} \quad RA = TC - TC(1+p)/(1+c),$$

will produce the same value of revenue adjustment regardless of the value estimated for c.

A-IV. DOES THE RAFP SIMPLE FORMULA MAKE SENSE?

The proposed RAFP Simple Formula may be expressed as follows:

> Revenue Adjustment equals the test period cost less the base period cost times the growth index.

While correct, this statement is not obvious on first reading. That being the case, it appears necessary to restate the RAFP Simple Formula in a more meaningful way. This alternate statement follows from equation (A18) which may be derived as follows:

$$\text{(A16)} \quad RA = TC - BC(1+g)$$

$$= TC - BC - gBC, \quad \text{or}$$

$$\text{(A18)} \quad RA = TC - (BC + GC)$$

Where: RA = revenue adjustment,

TC = test period cost (i.e., current costs),

BC = base period cost (i.e., old business costs),

g = output growth rate, and

GC = growth cost (i.e., new business costs).

Expressed simply, equation (A18) may be stated as follows:

> Revenue adjustment equals current costs less
> old business costs and new business costs

This statement makes sense since the revenue adjustment so derived would compensate the carriers for only those costs not recovered through either old or new business. Clearly, these costs must be recovered from a rate increase since there is no other feasible source of revenues.

It follows that the RAFP Simple Formula does make sense, and therefore should be accepted by the public at large.

A-V. RAFP ILLUSTRATION

To illustrate the various formulas developed in this paper consider the following situation:

Base Period

Revenue $= BR = \$85M$ (base period tariff),

Cost $= BC = \$85M$,

Test Period

Revenue $= TR = \$102M$ (base period tariff),

Cost $= TC = \$108M$, and

Rate of Inflation

Average $= c = 8\%$.

Given the above information we can solve the various formulas:

(A2) $$1+g = TR/BR = \$102M/\$85M = 1.20$$

(A4) $$BI = BR/BC = \$85M/\$85M = 1.00$$

(A5) $$TI = \frac{TR}{TC/(1+c)} = \frac{\$102M}{\$108M/1.08} = 1.02$$

(A6) $$1+p = TI/BI = 1.02/1.00 = 1.02$$

(A7) $$1+p = \frac{BC(1+g)(1+c)}{TC} = \frac{\$85M(1.20)(1.08)}{\$108M} = 1.02$$

(A8) $$IC = TC - TC/(1+c) = \$108M - \$108M/(1.08) = \$8M$$

(A9) $$RA = TC - TC\frac{(1+p)}{(1+c)} = \$108M - \$108M\frac{(1.02)}{(1.08)} = \$6M$$

(A11) $$PG = IC - RA = \$8M - \$6M = \$2M$$

(A12) $$PG = pTC/(1+c) = 2\% \times \$108M/1.08 = \$2M$$

(A13) $$GC = gBC = 20\% \times \$85M = \$17M$$

(A14) $$TC = BC + GC + IC - PG = \$85M + \$17M + \$8M - \$2M = \$108M$$

(A15) $$TC\frac{(1+p)}{(1+c)} = \$108M\frac{(1.02)}{(1.08)} = \$102M$$

$$BC(1+g) = \$85M(1.20) = \$102M$$

(A16) $TC = TC - BC(1+g) = \$108M - \$85M(1.20) = \$6M$

(A17) $\frac{1+p}{1+c} = \frac{1.02}{1.08} = 0.94444\ldots$

$\frac{BC}{TC}(1+g) = \frac{\$85M(1.20)}{\$108M} = 0.94444\ldots$

(A18) $RA = TC - (BC + GC) = \$108M - (\$85M + \$17M) = \$6M$

(A1) $TC = BC(1+g)\frac{(1+c)}{(1+p)} = \$85M(1.20)\frac{(1.08)}{(1.02)} = \$108M$

Note that if c where 12.5% rather than 8% we would obtain:

(A4) $BI = BR/BC = \$85M/\$85M = 1.00$,

(A5) $TI = \frac{TR}{TC/(1+c)} = \frac{\$102M}{\$108M/1.125} = 1.0625$,

(A6) $1+p = TI/BI = 1.0625/1.00 = 1.0625$, and

(A17) $\frac{1+p}{1+c} = \frac{1.0625}{1.1250} = 0.094444\ldots$ as before.

Thus, it is observed that the expression, (1+p)/(1+c), remains constant regardless of what value is attributed to the average rate of inflation, c.

A-VI. CONCLUSIONS FOR PART A

1. The following equations have been developed:

(A1) $TC = BC(1+g)(1+c)/(1+p)$

(A2) $1+g = TR/BR$

(A3) $PI = \frac{\text{Output}}{\text{Input}} = \frac{\text{Total Constant Dollar Revenue}}{\text{Total Constant Dollar Cost}}$

(A4) $BI = BR/BC$

(A5) $TI = \frac{TR}{TC/(1+c)}$

(A6) $1+p = TI/BI$

(A7) $1+p = \frac{BC}{TC}(1+g)(1+c)$

(A8) $IC = TC - TC/(1+c)$

(A9) $RA = TC - TC(1+p)/(1+c)$

(A10) $RA = IC - PG$

(A11) $PG = IC - RA$

(A12) $PG = pTC/(1+c)$

(A13) $GC = gBC$

(A14) $TC = BC + GC + IC - PG$

(A15) $TC(1+p)/(1+c) = BC(1+g)$

(A16) $RA = TC - BC(1+g)$

(A17) $\frac{1+p}{1+c} = \frac{BC}{TC}(1+g)$ = constant (given BC, TC and g)

(A18) $RA = TC - (BC + GC)$

Where: c = average rate of inflation,

g = rate of growth in business output,

p = rate of productivity gain,

1+c = inflation index,

1+g = growth index,

1+p = productivity gain index,

BC = base period costs,

BI = base period productivity index,

BR = base period revenues,

GC = growth cost,

IC = inflation cost,

PG = productivity gain,

PI = productivity index (constant dollar revenue/cost),

RA = revenue adjustment (value of rate adjustment),

TC = test period costs,

TI = test period productivity index, and

TR = test period revenues.

2. One may calculate the revenue adjustment by means of the following RAFP Simple Formula:

   Revenue adjustment equals the test period cost, less the base period cost times the growth index.

   The RAFP Simple Formula may also be restated:

   Revenue adjustment equals current cost less old business and new business cost.

3. The RAFP Simple Formula is the formula best suited for calculating revenue adjustment for the following reasons:

   (1) The formula adequately compensates the carriers for the uncontrollable increases in costs (due to inflation) offset by productivity gains.

   (2) The formula does not compensate for those costs associated with increased business volume.

   (3) The formula provides neither an incentive nor a disincentive to carrier efficiency, and therefore should not "interfere" with management decisions regarding the pattern of allocation of resources, debt/equity ratio, etc.

   (4) The formula is simple in structure, easy to execute, and avoids the need to determine rates of productivity gain and inflation which are subject to many technical difficulties.

   (5) The formula is defensible and in line with economic principles since it is grounded on the fundamental equation relating cost changes to growth, productivity gain and inflation, i.e.,

$$\begin{pmatrix}\text{The Test}\\ \text{Period Cost}\end{pmatrix} = \begin{pmatrix}\text{The Base}\\ \text{Period Cost}\end{pmatrix} \times \frac{(\text{Growth Index})(\text{Inflation Index})}{(\text{Productivity Gain Index})}$$

(6) The formula agrees with common sense as it only compensates for those costs not recoverable through old or new business revenues.

PART B: UNITS OF INPUT AND OUTPUT: ILLUSTRATION RE-EXAMINED

B-I. INTRODUCTION

In PART A, formulas were demonstrated by means of an illustration (p. 11). For a better appreciation of that illustration assume the following:

Input Unit = expense item (or part) costed at $1 in base period
Output Unit = service item (or part) priced at $1 in base period

Given the above convention the illustration may be restated:

Base Period

Revenue = 85M output units @ $1.00 = $85M = BR
Cost = 85M input units @ $1.00 = $85M = BC

Test Period

Revenue = 102M output units @ $1.00 = $102M = TR
Cost = 100M input units @ $1.08 = $108M = TC

B-II. ANALYSIS USING INPUT AND OUTPUT UNITS

1. The cost per input unit has increased from $1.00 to $1.08, indicating a rate of inflation of 8 cents per input unit, or c = 8% (as was given in the original illustration).

2. Since the number of input units used in the test period is 100M, then the increase in the cost of the input units represents an inflation cost (IC) of $8M (100M input units @ $0.08 each).

3. Production has increased from 1.00 output units per input unit (85M output units/85M input units) to 1.02 output units per input unit (102M output units/100M input units), indicating a rate of productivity gain of 2%.

4. If productivity had not improved, the required number of input units in the test period would have been 102M (to produce 102M output units at 1.00 output units per input unit), rather than the 100M input units actually used, indicating a saving of 2M input units (102M - 100M).

5. In terms of the base period cost of $1.00 per input unit, the 2M input units saved represents a productivity gain (PG) of $2M.

6. Or productivity gain (PG) is equal to the rate of productivity gain of 2% times the deflated test period cost of $100M (100M test period input units times base period unit price of $1.00), or $2M.

7. Neither the inflation cost (IC) nor the productivity gain (PG) includes the inflationary cost component of the saved input units (2M input units @ $0.08 = $0.16M). This is reasonable since this "component" is automatically cancelled, i.e., the "component" of productivity gain cancels the "component" of inflation cost. If this cancellation did not occur, IC and PG would be greater by $0.16M. I.E., the inflation cost would be $8.16M (102M input units @ $1.08), while the productivity gain would be $2.16M (2M units @ $1.08).

8. If the carrier had experienced a 20% change in scale in response to the 20% business growth, the increase in input units would have been 17M (20% of the 85M input units used in the base period).

9. In terms of the base period cost of $1.00 per input unit, this increase of 17M input units represents a growth cost (GC) of $17M.

10. The cost change from base period to test period (from $85M to $108M) may now be summarized as in Table B1:

	Term			Value
Base Period Cost	BC	85M units @ $1.00	=	$85M
+ Growth Cost @ 20%	+ GC	+17M units @ $1.00	=	+$17M
= Grown Base Period Cost	= BC(1+g)	102M units @ $1.00	=	$102M
- Productivity Gain @ 2%	- PG	-2M units @ $1.00	=	-$2M
= Deflated Test Period Cost	= TC/(1+c)	100M units @ $1.00	=	$100M
+ Inflation Cost @ 8%	+ IC	100M units @ +$.08	=	+$8M
= Test Period Cost	= TC	100M units @ $1.08	=	$108M

Table B1: Cost Change Summary (8% Inflation)

11. Since the inflation cost (IC) is $8M, while the productivity gain (PG) is $2M, then offsetting one with the other gives a revenue adjustment (RA) of $6M ($8M-$2M).

12. Or the revenue adjustment may be obtained by subtracting "the test period cost in the absence of productivity gain and inflation" of $102M (100M x 1.02 units @ $1.08 / 1.08) from the test period cost of $108M (100M units @ $1.08) to get $6M ($108M - $102M) as before.

13. Also revenue adjustment may be obtained by subtracting the "grown base period cost" of $102M (120% of 85M units @ $1.00) from the test period cost of $108M (100M units @ $1.08) to get $6M ($108M-$102M).

14. Note, with reference to Table B1, that the "grown base period cost" of $102M is identical to the "test period cost in the absence of productivity gain and inflation" of $102M, which explains the similarities of observations 12 and 13, i.e.,

Since BC(1+g) = TC(1+c)/(1+p) = $102M (Table B1),
and RA = TC - TC(1+c)/(1+p) = $108M - $102M = $6M (Obs. 12.),
then RA = TC - BC(1+g) = $108M - $102M = $6M (Obs. 13.).

15. Finally, the revenue adjustment (RA) may be obtained by taking the total cost increase of $23M ($108M - $85M) and subtracting the growth cost (GC) of $17M (17M units @ $1.00) to get $6M ($23M - $17M).

16. If the rate of inflation were in fact 12.5% rather than 8%, then the cost per input unit in the test period would be $1.125 rather than $1.08, and applied to 96M input units ($108M/$1.125) rather than 100M ($108/$1.08), leading to the cost changes shown in Table B2:

	Term			Value
Base Period Cost	BC	85M units @ $1.00	=	$85M
+ Growth Cost @ 20%	+ GC	+17M units @ $1.00	=	+$17M
= Grown Base Period Cost	= BC(1+g)	102M units @ $1.00	=	$102M
- Productivity Gain @ 6.25%	- PG	-6M units @ $1.00	=	-$6M
= Deflated Test Period Cost	= TC/(1+c)	96M units @ $1.00	=	$96M
+ Inflation Cost @ 12.5%	+ IC	96M units @ +$.125	=	+$12M
= Test Period Cost	= TC	96M units @ $1.125	=	$108M

Table B2: Cost Change Summary (12.5% Inflation)

Thus, 1+g = 102M units/85M units = 1.20, or g = 20%,
1+p = 102M units/96M units = 1.0625, or p = 6.25%,
and 1+c = $1.125/$1.00 = 1.125, or c = 12.5%.

Therefore, (1+p)/(1+c) = 1.0625/1.125 = 0.9444... as before.

Also, RA = IC - PG = $12M - $6M = $6M as before,
RA = TC - TC(1+c)/(1+p) = $108M - $102M = $6M as before,
and RA = TC - BC(1+g) = $108M - $102M = $6M as before.

One observes that a change in the estimated rate of inflation has no effect on the calculated value of revenue adjustment.

B-III. ANALYSIS OF A THREE SERVICE SITUATION

Having analysed the RAFP Illustration as above, one may well enquire regarding the nature of a similar analysis involving a multiservice company with each service having different rates of growth and inflation.

In reply, consider three services having costs, rates of growth, and rates of inflation as shown in Table B3:

Service Number	i	1	2	3	Total
Base Period Cost	BC_i	$10.0M	$50.0M	$25.0M	$85.0M
Rate of Growth	g_i	30%	10%	36%	20%
Rate of Inflation	c_i	12%	8%	6.6%	8%
Test Period Cost	TC_i	$14.0M	$58.3M	$35.7M	$108.0M

Table B3: Three Service Company

Using the above convention, i.e., one input unit equals an expense item or part costed at $1.00 in the base period, we may analyse the data on the three services and arrive at the following three tables of cost changes (Tables B4, B5 and B6):

Service 1	Term		Value
Base Period Cost	BC_1	10.0M units @ $1.00 =	$10.0M
+ Growth Cost @ 30%	$+ GC_1$	+3.0M units @ $1.00 =	+$3.0M
= Grown Base Period Cost	$= BC_1(1+g_1)$	13.0M units @ $1.00 =	$13.0M
- Productivity Gain @ 4%	$- PG_1$	-0.5M units @ $1.00 =	-$0.5M
= Deflated Test Period Cost	$= TC_1/(1+c_1)$	12.5M units @ $1.00 =	$12.5M
+ Inflation Cost @ 12%	$+ IC_1$	12.5M units @ +$.12 =	+$1.5M
= Test Period Cost	$= TC_1$	12.5M units @ $1.12 =	$14.0M

Table B4: Service 1 Cost Changes

Service 2	Term		Value
Base Period Cost	BC_2	50.0M units @ $1.00 =	$50.0M
+ Growth Cost @ 10%	$+ GC_2$	+5.0M units @ $1.00 =	+$5.0M
= Grown Base Period Cost	$= BC_2(1+g_2)$	55.0M units @ $1.00 =	$55.0M
- Productivity Gain @ 1.85%	$- PG_2$	-1.0M units @ $1.00 =	-$1.0M
= Deflated Test Period Cost	$= TC_2/(1+c_2)$	54.0M units @ $1.00 =	$54.0M
+ Inflation Cost @ 8%	$+ IC_2$	54.0M units @ +$.08 =	+$4.3M
= Test Period Cost	$= TC_2$	54.0M units @ $1.08 =	$58.3M

Table B5: Service 2 Cost Changes

Service 3	Term		Value
Base Period Cost	BC_3	25.0M units @ $1.00 =	$25.0M
+ Growth Cost @ 36%	$+ GC_3$	+9.0M units @ $1.00 =	+$9.0M
= Grown Base Period Cost	$= BC_3(1+g_3)$	34.0M units @ $1.00 =	$34.0M
- Productivity Gain @ 1.5%	$- PG_3$	-0.5M units @ $1.00 =	-$0.5M
= Deflated Test Period Cost	$= TC_3/(1+c_3)$	33.5M units @ $1.00 =	$33.5M
+ Inflation Cost @ 6.6%	$+ IC_3$	33.5M units @ +$.066=	+$2.2M
= Test Period Cost	$= TC_3$	33.5M units @ $1.066=	$35.7M

Table B6: Service 3 Cost Changes

The values in the above three tables may be summed as shown in Table B7 on the following page.

(Note that overall growth and overall inflation have been made, by design, to agree with the original illustration (Table B1, p. 18). However, the agreement with respect to overall productivity gain follows naturally from the consistency of the analytical method used.)

Service 1, 2 and 3	Term				Value
Service Number		1	2	3	Total
Base Period Cost	BC	$10.0M	$50.0M	$25.0M	$85.0M
+ Growth Cost @ 20%	+ GC	+$3.0M	+$5.0M	+$9.0M	+$17.0M
= Grown Base Period Cost	= BC(1+g)	$13.0M	$55.0M	$34.0M	$102.0M
- Productivity Gain @ 2%	- PG	-$0.5M	-$1.0M	-$0.5M	-$2.0M
= Deflated Test Period Cost	= TC/(1+c)	$12.5M	$54.0M	$33.5M	$100.0M
+ Inflation Cost @ 8%	+ IC	+$1.5M	+$4.3M	+$2.2M	+$8.0M
= Test Period Cost	= TC	$14.0M	$58.3M	$35.7M	$108.0M

Table B7: Total Cost Changes By Summation

With regard to revenue adjustments per service, both individually and in total, these may be obtained using the formula, RA = IC - PG (Table B8):

Service 1, 2 and 3	Term				Value
Service Number		1	2	3	Total
Inflation Cost	IC	$1.5M	$4.3M	$2.2M	$8.0M
- Productivity Gain	- PG	-$0.5M	-$1.0M	-$0.5M	-$2.0M
= Revenue Adjustment	= RA	$1.0M	$3.3M	$1.7M	$6.0M

Table B8: Revenue Adjustment (CTC Formula)

The revenue adjustments may also be obtained using the RAFP Simple Formula, RA = TC - BC(1+g), as shown in Table B9:

Service 1, 2 and 3	Term				Value
Service Number		1	2	3	Total
Test Period Cost	TC	$14.0M	$58.3M	$35.7M	$108.0M
- Grown Base Period Cost	- BC(1+g)	-$13.0M	-$55.0M	-$34.0M	-$102.0M
= Revenue Adjustment	= RA	$1.0M	$3.3M	$1.7M	$6.0M

Table B9: Revenue Adjustment (RAFP Simple Formula)

Finally, to demonstrate the overall consistency of the above method of analysis, we may apply the fundamental equation relating cost changes to growth, productivity gain and inflation, i.e.,

$$\begin{pmatrix}\text{The Test}\\ \text{Period Cost}\end{pmatrix} = \begin{pmatrix}\text{The Base}\\ \text{Period Cost}\end{pmatrix} \times \frac{\text{(Growth Index) (Inflation Index)}}{\text{(Productivity Gain Index)}},$$

to the three services individually and in total, as shown in Table B10:

Service 1, 2 and 3	Term				Value
Service		1	2	3	Total
Base Period Cost	BC	$10.0M	$50.0M	$25.0M	$85.0M
x Growth Index	x (1+g)	1.30	1.10	1.36	x 1.20
/ Productivity Gain Index	/ (1+p)	1.04	1.0185	1.015	/ 1.02
x Inflation Index	x (1+c)	1.12	1.08	1.066	x 1.08
= Test Period Cost	= TC	$14.0M	$58.3M	$35.7M	$108.0M

Table B10: Fundamental Economic Formula

B-IV. REVENUE ADJUSTMENTS FOR COMPETITIVE AND MONOPOLY SERVICES

Considering the THREE SERVICE SITUATION (Part B-V), it should be clear that revenue adjustments could be calculated for groups of monopoly and competitive services entirely from a knowledge of costs and growth rates. For example, if Service 1 were a competitive service while Services 2 and 3 were monopoly services, then from Table B3 we could construct Table B11:

Service Group		Compet	Monop	Total
Base Period Cost	BC	$10M	$75M	$85M
Rate of Growth	g	30%	18.7%	20%
Test Period Cost	TC	$14M	$94M	$108M

Table B11: Competitive And Monopoly Situation

From Table B11 we may easily construct Table B12 showing the appropriate revenue adjustments:

Service Group		Compet	Monop	Total
Test Period Cost	TC	$14M	$94M	$108M
- Grown Base Period Cost	- BC(1+g)	-$13M	-$89M	-$102M
= Revenue Adjustment	= RA	$1M	$5M	$6M

Table B12: Revenue Adjustment (Competitive And Monopoly)

The above indicates that the distributions of revenue adjustment between competitive and monopoly subscribers should be $1M and $5M respectively.

As a common-sense test, first consider the competitive group with a cost increase of $4M ($14M-$10M). Of this, $3M should be recovered by the 30% growth (base period competitive revenues of $10M times 30%), with $1M remaining. Now consider the monopoly group with a cost increase of $19M ($94M-$75M). Of this, $14M should be recovered by the 18.7% growth (base period monopoly revenues of $75M times 18.7%), with $5M remaining.

Clearly, these remaining costs should be recovered by rate increases in competitive and monopoly services amounting to $1M and $5M respectively, i.e., amounts equal to the revenue adjustments calculated for Table B12 using the RAFP Simple Formula.

B-V. CONCLUSIONS FOR PART B

1. A deeper understanding of the RAFP Illustration (and the RAFP Simple Formula) may be obtained by adopting the following convention:

Input Unit = expense item (or part) costed at $1 in base period,

Output Unit = service item (or part) priced at $1 in base period.

2. By adopting the above convention, and defining Test Period Revenues as <u>test period revenues adjusted to the base period tariff</u>, the following analytical steps may be taken (See Table B1, p. 18, for summary):

(1) Base Period Output Volume = Base Period Revenues / $1.00

e.g., $85M / $1.00 = 85M output units

(2) Test Period Output Volume = Test Period Revenues / $1.00

e.g., $102M / $1.00 = 102M output units

(3) Growth Index = $\frac{\text{Test Period Output Volume}}{\text{Base Period Output Volume}}$

e.g., 102M / 85M = 1.20

(4) Growth Rate = Growth Index - 1

e.g., 1.20 - 1 = 20%

(5) Base Period Input Volume = Base Period Cost / $1.00

e.g., $85M / $1.00 = 85M input units

(6) Growth Cost = Growth Rate x Base Period Input Volume x $1.00

e.g., 20% x 85M x $1.00 = $17M

(7) Grown Base Period Cost = Base Period Cost x Growth Index

e.g., $85M x 1.20 = $102M

(8) Test Period Unit Cost = Inflation Index x $1.00

e.g., 1.08 x $1.00 = $1.08 per input unit

(9) Test Period Unit Inflation Cost = Test Period Unit Cost - $1.00

e.g., $1.08 - $1.00 = $0.08 per input unit

(10) Test Period Input Volume = $\frac{\text{Test Period Cost}}{\text{Test Period Unit Cost}}$

e.g., $108M / $1.08 = 100M input units

(11) Inflation Cost = Test Period Input Volume x Test Period Unit Inflation Cost

e.g., 100M x \$0.08 = \$8M

(12) Deflated Test Period Cost = $\frac{\text{Test Period Cost}}{\text{Inflation Index}}$

e.g., \$108M / 1.08 = \$100M

(13) Productivity Gain = Grown Base Period Cost - Deflated Test Period Cost

e.g., \$102M - \$100M = \$2M

(14) Productivity Gain Index = $\frac{\text{Grown Base Period Cost}}{\text{Deflated Test Period cost}}$

e.g., \$102M / \$100M = 1.02

(15) Productivity Gain Rate = Productivity Gain Index - 1

e.g., 1.02 - 1 = 2%

(16) Revenue Adjustment = Inflation Cost - Productivity Gain

e.g., \$8M - \$2 = \$6M

(17) Revenue Adjustment = Test Period Cost - Grown Base Period Cost

e.g., \$108M - \$102M = \$6M

3. The above analysis may be performed just as well on individual services as on the company as a whole while retaining overall consistency of results. For example see Tables B7 through B10.

4. The ease with which the above type of analysis may be performed, and the understanding it conveys should encourage the general public to accept the RAFP Simple Formula.

PART C: - RAFP AND THE ANTI-INFLATION ACT REGULATIONS

C-I. INTRODUCTION AND ANALYSIS

The purpose of a Rate Adjustment Formula Procedure (RAFP) is to determine an allowable Revenue Adjustment: the dollar value of the rate adjustment applicable in the period following the test period. Since this is the same objective as the Anti-Inflation Board (AIB), when applying the Anti-Inflation Act Regulations, it is of interest to compare the two methods.

In the paper entitiled "Attack On Inflation", a policy statement tabled in the House of Commons by the Honourable Donald S. Macdonald, Minister of Finance, 14 October 1975, the basic price control guidelines are stated:

> The general principle is that increases in prices should be limited to amounts no more than required to cover net increases in costs.[14]
>
> Firms which are able to allocate costs to individual products are expected to increase prices of these products by no more than increases in costs allocated to this product.[15]

For a firm with n products, the above statements imply that the change in price for product i should equal the change in unit cost of that product:

$$\text{(C1)} \quad \begin{pmatrix}\text{Change In}\\ \text{Unit Price}\end{pmatrix}_i = \begin{pmatrix}\text{Change In}\\ \text{Unit Cost}\end{pmatrix}_i$$

This implies a change in total revenue (Revenue Adjustment) for product i calculated as follows:

$$\text{(C2)} \quad \begin{pmatrix}\text{Revenue}\\ \text{Adjustement}\end{pmatrix}_i = \begin{pmatrix}\text{Change In}\\ \text{Unit Cost}\end{pmatrix}_i \times \begin{pmatrix}\text{Test Period}\\ \text{Volume}\end{pmatrix}_i$$

Equation (C2) may be restated in a number of alternate ways:

(C3) $$\begin{pmatrix}\text{Revenue}\\\text{Adjustment}\end{pmatrix}_i = \begin{pmatrix}\text{Test Period} - \text{Base Period}\\\text{Unit Cost} \quad\quad \text{Unit Cost}\end{pmatrix}_i \times \begin{pmatrix}\text{Test Period}\\\text{Volume}\end{pmatrix}_i$$

or

(C4) $$\begin{pmatrix}\text{Revenue}\\\text{Adjustment}\end{pmatrix}_i = \begin{pmatrix}\dfrac{\text{Test Period Cost}}{\text{Test Period Volume}} - \dfrac{\text{Base Period Cost}}{\text{Base Period Volume}}\end{pmatrix}_i \times \begin{pmatrix}\text{Test Period}\\\text{Volume}\end{pmatrix}_i$$

or

(C5) $$\begin{pmatrix}\text{Revenue}\\\text{Adjustment}\end{pmatrix}_i = \begin{pmatrix}\text{Test Period}\\\text{Cost}\end{pmatrix}_i - \begin{pmatrix}\text{Base Period}\\\text{Cost}\end{pmatrix}_i \times \begin{pmatrix}\dfrac{\text{Test Period Volume}}{\text{Base Period Volume}}\end{pmatrix}_i$$

or

(C6) $$\begin{pmatrix}\text{Revenue}\\\text{Adjustment}\end{pmatrix}_i = \begin{pmatrix}\text{Test Period}\\\text{Cost}\end{pmatrix}_i - \begin{pmatrix}\text{Base Period}\\\text{Cost}\end{pmatrix}_i \times \begin{pmatrix}\text{Growth}\\\text{Factor}\end{pmatrix}_i$$

or

(C7) $$RA_i = TC_i - BC_i(1+g_i)$$

For the firm as a whole, all n products must be taken into account. Thus we get the following:

(C8) $$\sum_{i=1}^{n} RA_i = \sum_{i=1}^{n} TC_i - \sum_{i=1}^{n} BC_i(1+g_i)$$

or

(C9) $$RA = TC - \sum_{i=1}^{n} BC_i - \sum_{i=1}^{n} BC_i \cdot g_i$$

or

(C10) $$RA = TC - BC - BC\left(\frac{\sum_{i=1}^{n} BC_i \cdot g_i}{\sum_{i=1}^{n} BC_i}\right)$$

or

(C11) $$RA = TC - BC - BCg$$

or

(C12) $$RA = TC - BC(1+g)$$

$$\text{Where: } RA = \sum_{i=1}^{n} RA_i = \text{ total revenue adjustment,}$$

$$TC = \sum_{i=1}^{n} TC_i = \text{ total test period cost,}$$

$$BC = \sum_{i=1}^{n} BC_i = \text{ total base period cost, and}$$

$$g = \left(\frac{\sum_{i=1}^{n} BC_i \cdot g_i}{\sum_{i=1}^{n} BC_i} \right) = \text{ average growth rate.}$$

Equation (C12) is the same RAFP Simple Formula proposed earlier. Thus, the RAFP Simple Formula conforms to the Anti-Inflation Act Regulations, and therefore should be acceptable to either the AIB or its successor.

C-II. AIB ILLUSTRATION

To demonstrate the compatibility of the RAFP Simple Formula with the Anti-Inflation Act Regulations, consider Company A which has three products (or services) having costs and volumes given in Table C1:

	BASE PERIOD		TEST PERIOD	
Product	Output Volume (units)	Output Unit Cost	Output Volume (units)	Output Unit Cost
1	10,000	$1,000	13,000	$1,077
2	25,000	2,000	27,500	2,120
3	5,000	5,000	6,800	5,250

Table C1: Three Product Situation

Table C1 indicates that the unit costs of products 1, 2 and 3 have increased (due to a combination of inflation and productivity gain) by $77, $120 and $250 respectively, leading to a total required revenue adjustment in the test year of $6.0M as shown in Table C2:

Product	Test Period Output Volume (units)	Output Unit Cost Increase	Product Revenue Adjustment RA_i
1	13,000	$77	$1.0M
2	27,500	120	3.3
3	6,800	250	1.7
Revenue Adjustment (RA)			$6.0M

Table C2: Revenue Adjustment Per Guidelines

Under the price control quidelines, Company A would be permitted to increase its prices to just recover the $6M. Significantly, this $6M may also be calculated using the RAFP Simple Formula: First, we determine the total base period cost (BC) to be $85M as shown in Table C3:

	BASE PERIOD		
Product	Output Volume (units)	Output Unit Cost	Product Cost BC_i
1	10,000	$1,000	$10.0M
2	25,000	2,000	50.0
3	5,000	5,000	25.0
Base Period Cost (BC)			$85.0M

Table C3: Base Period Cost

Secondly, we determine the total test period cost (TC) to be $108M as shown in Table C4:

	TEST PERIOD		
Product	Output Volume (units)	Output Unit Cost	Product Cost TC_i
1	13,000	$1,077	$14.0M
2	27,500	2,120	58.3
3	6,800	5,250	35.7
Test Period Cost (TC)			$108.0M

Table C4: Test Period Cost

Thirdly, we determine the total growth cost (GC) to be \$17M as shown in Table C5:

Product	BASE Period Output Volume (units)	TEST Period Output Volume (units)	Output Unit Growth Rate g_i	BASE Period Product Cost BC_i	Product Growth Cost $BC_i \cdot g_i$
1	10,000	13,000	30%	\$10.0M	\$3.0M
2	25,000	27,500	10%	50.0	5.0
3	5,000	6,800	36%	25.0	9.0
Growth Cost (GC)					\$17.0M

Table C5: Growth Cost

Fourthly, we determine the average growth rate (g) to be 20% by dividing the growth cost by the base period cost:

$$g = \frac{\sum_{i=1}^{n} BC_i \cdot g_i}{\sum_{i=1}^{n} BC_i} = \frac{GC}{BC} = \frac{\$17M}{\$85M} = 20\%$$

Finally, we substitute the appropriate values into the RAFP Simple Formula to obtain the required Revenue Adjustment (RA):

$$\text{(C12)} \qquad RA = TC - BC(1+g)$$
$$= \$108M - \$85M(120\%) = \$6M \text{ as before.}$$

In the actual application of the RAFP Simple Formula, values for the test period costs (TC) and base period costs (BC) could be easily obtained from the income statement. However, the average growth rate (g) would be difficult to obtain using the above formula. To overcome this difficulty an alternate procedure for determining growth is proposed:

First, we assume that items are sold at cost in the base period (profits are treated as a cost), and determine the base period revenue (BR) to be $85M as shown in Table C6:

	BASE PERIOD		
Product	Volume (units)	Prices	Revenues
1	10,000	$1,000	$10M
2	25,000	2,000	50
3	5,000	5,000	25
Base Period Revenues (BR)			$85M

Table C6: Base Period Revenues

Secondly, we price the test period volumes in terms of the base period prices to obtain a test period adjusted revenue (TR) of $102M as shown in Table C7:

Product	TEST Period Volume (units)	BASE Period Prices	TEST Period Adjusted Revenues
1	13,000	$1,000	$13M
2	27,500	2,000	55
3	6,800	5,000	34
Test Period Adjusted Revenues (TR)			$102M

Table C7: Test Period Adjusted Revenues

Finally, we determine the growth factor (1+g) by dividing the Base Period Revenues (BR) into the Test Period Adjusted Revenues (TR) to get 1.20 ($102M/$85M), or a rate of growth of 20% as before.

(Note that the illustration shown here is essentially the same as that analysed previously in Part B. This is intentional to permit comparison of the analytical methods used.)

C-III CONCLUSIONS FOR PART C

1. The RAFP Simple Formula is consistent with the price control guidelines of the Anti-Inflation Act Regulations since both methods produce the same Revenue Adjustment: the value of the rate adjustment permitted in the period following the test period (equal to the extra revenue that would have made the carrier as well off in the test period as it had been in the base period).

2. The RAFP Simple Formula is the easiest way to determine the Revenue Adjustment as only four values are required:

   (1) Total Base Period Cost (BC),
   (2) Total Test Period Cost (TC),
   (3) Total Base Period Revenues (BR), and
   (4) Total Test Period Adjusted Revenues (TR).

   Of these values, the first three are immediately available from the income statement, while the fourth is easily obtained by pricing the test period output at base period tariff. Using these four values, the Revenue Adjustment (RA) is calculated as follows:

$$\begin{aligned} RA &= TC - BC(1+g) \\ &= TC - BC(TR/BR) \end{aligned}$$

3. Considering that the RAFP Simple Formula is easy to apply, easy to audit and conforms to the price control guidelines, the AIB or its successor should endorse its use in a Rate Adjustment Formula Procedure.

COMMENT ON

"GLOBAL FACTOR PRODUCTIVITY (GFP) AND EDF'S MANAGEMENT"

RON MILLEN

This paper outlines the calculation, results and use of global factor productivity at EDF. The paper is concise and easily understood but there are a few instances where further elaboration or clarification might be useful.

Page 4 - It appears that the productivity index used is a chained Laspeyres type rather than a Törnqvist Divisia index. Is this true?

Page 6 - The social rate of return used is 9%, while the actual capital charges and operating result is 5-6% in 1977 and 1978 (based on Table 1). There is no comment on what effect this higher weight on the capital input would have on calculated productivity.

Page 10 - The comments on historic changes in the productivity index are rather spurious and do not relate productivity changes to underlying factors which management might associate with particular years. The chart of productivity gains appears to show three separate periods:

1961-66 - decreasing output and productivity growth
1967-72 - increasing output and productivity growth
1973-78 - decreasing output and productivity growth.

Certain years appear to be slightly off trend when considering the correspondence of output and productivity growth:

1962,63,66,69,76
- productivity growth lower than expected
1975 - productivity growth higher than expected.

These trends and particular yearly results should be explained in terms of underlying causal factors if management is to find this a useful tool.

Page 16 - As pointed out in the paper, several examples of increasing scale exist. In spite of this it may well be that decreasing returns predominate as the company is forced to move from hydro to thermal to nuclear technologies. I believe this has been the experience in Canada where we initially had an abundance of cheap hydro power. Is it possible to calculate disaggregate productivity estimates for each type of generation source?

Page 16 - The problem of factor weights is particularly severe in this case where fuel prices have changed so rapidly. Factor prices were held constant for capital and this technique might also be used for fuel.

General Comment - The author provides little evidence of exactly how productivity is used in internal budgeting, planning processes, etc. This type of discussion is implied by the title.

COMMENT ON "NET INCOME AND PRODUCTIVITY ANALYSIS (NIPA) AS A PLANNING MODEL"

RON MILLEN

Calculation of the dollar value of productivity gains instead of the percentage increase in a productivity index aids in making the measure and its implications more meaningful to managers. A.T.&T. first proposed this procedure more than ten years ago in a handbook on productivity. The framework has been elaborated since that time but the concept is the same. Of course, the dollar value of productivity gains can be calculated either for historic data or for a forecast scenario. As a planning model, the proposed NIPA model is somewhat limited in its present form since all of the output and input variables are exogenous and must be specified in advance. The model simply calculates the dollar value of productivity gains and relates these to other dollar values implied in the forecast. On page 30, the authors point out that they plan to extend the model to include demand conditions, operating constraints and target variables. This should improve the usefulness of the model for planning purposes.

COMMENT ON "TOTAL FACTOR PRODUCTIVITY FOR MANAGEMENT: THE POST-MORTEM AND PLANNING FRAMEWORKS"

RON MILLEN

The first section of this paper extends the NIPA analysis by including a residual term that is the repository of all deviations from planned levels for each of the other variables. The authors are merely suggesting that actual versus budgeted results can be compared using the NIPA accounting identity as well as the standard income statement identity.

A second section of the paper proposes a more active approach to budgeting. A preliminary top-down budget is set given demand forecasts, rate of return requirements and desired productivity growth. The authors assume that a reliable cost function for the firm can be estimated. They also assume constant returns to scale for this function. Unfortunately, the validity of this premise remains to be established.

COMMENT ON THE MANAGEMENT APPLICATIONS OF PRODUCTIVITY ANALYSIS

R.E. OLLEY

The question of management applications of productivity analysis is very much in its early stages of exploration. Productivity measures provide firm or industry wide data which are every bit as important socially as profit is privately, as a measure of performance. As with the profit measure, the productivity measure is very general, embracing the whole firm's performance, or that of large segments of it. For that reason, it is not apparent from a historical series in respect to either measure, taken by itself without supplementary analyses, just what managers should do specifically to increase performance by the measure. Profit has been measured for a very long time and arouses immediate and urgent interest because of its implications for survival. Therefore, well understood, if very complex, analyses have been developed and continue to develop, to permit managers to work toward increased profit. That is, the generalization about a firm's performance, labelled profit, can be factored into specific and actionable insights for managerial application. Productivity, on the other hand, is relatively very new as a measurable concept for the whole firm. (It is not particularly new in its various partial forms such as operations per hour, sales per square foot, miles per gallon, and thousands more such micro measures.) It is not as yet clear how the measure of productivity can be analysed in such manners as to provide operationally useful managerial insight.

It is therefore not surprising that all of these papers are somewhat tentative in their procedures. All wrestle with problems which deserve the effort, and pursue avenues which have to be pursued sooner or later. Seen in this way, all four papers are valuable early probes into the area of managerial and regulatory applications of productivity analyses.

Both the Reimeringer and Chaudry-Burnside papers analyse the financial benefits to which observed productivity gains equate, then estimate how those benefits were distributed or spent on additional input costs, profit, and other items. The analyses conform to the exigencies of the product exhaustion theorem, with profit functioning, of course, as the residual category. This is a valuable exercise.

The Reimeringer paper makes, but does not elaborate upon, the extremely important point that there is a relationship between measured productivity gain and the quality of service. *Ceteris paribus*, that relationship in inverse, that is, lowered quality will lead to higher measured total or total factor productivity and vice versa, in most cases. This observation should constitute a very important warning against the uncritical establishment of productivity targets. At the very least it can be suggested that any unrealism in the targets will become the cause of quality variations as managers set out to meet those targets. Electricité de France, upon whose experience the Reimeringer paper is based, has attempted to set productivity targets. It would be useful to know what the experience was with that practice and whether it was useful enough to continue.

Chaudry and Burnside, basing their analysis on work which has presumably been carried out at AT & T, develop their NIPA model as a method of showing where the financial benefits from productivity gains (and other income augmenting factors) were spent or used up. The purpose is to provide managers with some insight as to the actions which must be taken to remedy profit shortfalls. This is a laudable purpose but the question which remains unanswered is what managers should do and how the NIPA model helps them to make better decisions. Profit shortfalls in a regulated utility can be met by price increases; that mechanism will always work since regulated utilities presumably have unexploited monopoly or oligopoly power which is not available for use because of regulation. One does not need the NIPA model to make that observation.

But what else can managers find to do that would not be obvious with the application of a little common sense? What can they do to remedy profit shortfalls if competitive entry becomes more significant? These questions remain unanswered. Intuitively one suspects that there are answers to such questions. The component series going into productivity measurement provide rich mines of data on labour, capital, and material inputs. Can these data be used to help inform strategic managerial planning? One suspects that the answer must be affirmative. It would be useful to know if there is any experience with the application of productivity data for labour force planning, construction program sizing, technology planning, and other input related managerial initiatives. Such uses would begin to build the required bridges between macro measurements of productivity and the thousands of micro efficiency measures with which telecommunications company managers are routinely familiar. Just as such bridges exist to permit individual firm activities to be related to the macro measure of profit, so too must they come to exist between efficiency measures and productivity measures. Otherwise there is little likelihood that productivity measures, however extensively they may be manipulated, will provide much more managerial insight than is already readily available to managers from other sources such as engineering and accounting models.

On the output side of a firm's activity parallel questions may be raised. If output mix is part of the explanation of productivity gains, how can that observation be made operationally useful for marketing, product or service development, or other market strategies? Similarly, if economies of scale (or scope) exist, what operational consequences do they have for future actions by the firm? It may be premature to ask such questions, certainly of the Reimeringer or Chaudry-Burnside papers, but they are potentially important ones for this area of study.

Denny, De Fontenay, and Werner (DFW) develop another form of NIPA model, called UNIPA. Generically it is similar to the previous one but it is extended in an interesting manner. It becomes effectively a generalized planning model for the firm, subject to all of the strengths and weaknesses of such models. What is unique here is that this approach to planning places emphasis on the real (deflated) variables. It is these which managers manage in their day-to-day activities. It is, furthermore, these which set in train the financial events which become profit (or loss) for the firm. Thus, the DFW approach promises the eventual capability to go directly from productivity measures and analyses to operationally useful observations about how the firm's profit and productivity may be improved; that progression, while complex, appears to be capable of being made comprehensible to non-economists, of whom management is principally composed. It is too early to ask that the paper's promise in this regard be fulfilled, thus it is no criticism to say that it is not. What is important is that the direction and emphasis, described during the verbal presentation as providing the basis for rational expectations about what can be achieved, appears to be capable of being made fruitful. In all probability the Reimeringer and Chaudry-Burnside approaches will emerge as subsystems within the DFW approach, as nearly as one can now intuit the outcome of this line of analysis.

The Goodier paper addresses an extremely important question. That is, whether there is any mechanism which can be developed to enable regulators to permit rate increases to regulated utilities when inflationary forces are rampant, making rate applications frequent and rate relief in some sense obviously necessary.

Mr. Goodier correctly observes that forecasts of productivity gain are difficult to obtain and subject to both argument when they are calculated and error as history turns them into actualities. Only a little less contentious is the problem of estimating the specific impact of inflationary conditions as they bear on the firm and make rate increases necessary. Mr. Goodier assumes that real growth is readily measurable. This is true for firms with short product lists and infrequent introductions of new products, since ordinary demand forecasts contain a measure of growth which is easy to obtain. Where the firm makes frequent new product introductions, when output composition changes, or where individual prices are changed as a result of special hearings outside of general rate applications, the measure of real growth becomes more difficult to obtain. It too becomes contentious. Measurement questions are resolved only as part of an exercise much like the general process involved in productivity measurement with its associated price indexes and weighting patterns.

Mr. Goodier's analysis commences with the observation (Equation A1) that

$$\text{Total costs in any test year} = \frac{\text{Total costs in the base or comparison year, times one plus the rate of inflation}}{\text{One plus the rate of productivity growth}}$$

By simple algebraic manipulation he arrives at the conclusion (Equation A16) that

Revenue adjustment = Total Test Period Costs minus base period costs times one plus the growth rate in real output

This conclusion is a tautology at all times and one with a unique value once total test period costs are established and a target rate of return is known. That being so, it is easy enough to say that productivity gain and rate of inflation may be dropped from the calculation since they must have a ratio to one another which is a unique value once total costs are established.

To proceed to state that the tautology is a valid device to permit a ready calculation of what the revenue adjustment to be allowed by any regulator should be is to ignore the real problems which face regulators in this regard.

It is unclear whether Mr. Goodier envisages the test year for revenue adjustment as an historic year or a prospective one. If it is an historic year then total costs have to be tested for validity. Otherwise any cost increases, within broad limits, will be allowed as valid. This amounts to little more than a blatant cost-plus contract for the regulated utility, being so because there is no ready way to determine the validity of the total costs, short of a full scale rate hearing. If the test year is prospective then the same observation is even more true since all costs are forecast; any breakdown of costs into price and quantity terms would be both too detailed to be meaningful if it were carried to analytically meaningful lengths, and fraught with even more complexities than are measurements of productivity gains. Again, nothing short of a full scale hearing could adequately test the multi-dimensional aspects of cost. In short, Mr. Goodier's equation A16 implicitly assumes away the problems with which regulators must, in applied regulation, concern themselves. To style that assumption a solution to regulatory burden is simply completely wrong.

Regulators, if their hand in the process is to mean anything substantial, have to question total costs directly. To do that they have to test that those costs have been minimized at current expected grades of service, under current general economic conditions, and with available production techniques. Effectively that questioning, however it is carried out in particular detail, amounts to assessing whether productivity gains have been reasonable and whether the impact of broad inflationary conditions (or more generally any pattern of current price conditions) has been minimized by management in its operations of the particular firm. No responsible regulator can avoid that burden. To shoulder it the regulator can either assess the expected productivity gains and expected specific impacts of inflation by the traditional methods of detailed examination of operating numbers, or establish expected values for overall productivity gain and inflationary impacts, testing the proposed revenue requirement with those two values. Assuming away the problem does not solve it.

Complex as the handling of expected values for productivity gain and specific impact of inflation may be, it is only by this route that rate hearings can be abbreviated or obviated at least for some years. There are procedures by which productivity gains can be predicted, not accurately but closely enough to prevent much consequential error, at least from a social point of view. At least two such methods are the following. One is to analyse the implied productivity gain in the pro forma budget, then test it against historic experience. This amounts to a forecast of productivity gain. While never perfectly accurate, these forecasts tend to be a bit too high but not to be grossly in error. The chances of the firm receiving a windfall gain are small and of its receiving a significant gain are negligible, particularly when it is recognized that the regulator does not have to wait a whole year to learn the profit implications of the new rates, these being available at least quarterly. A second method of setting the productivity figure is to project it from historical experience on the basis of expected final demand aggregates and other broad variables. This method produces less reliable forecasts than the first but the error is still not large. It does tend to be random about the mean realized value which increases somewhat the chances of windfall gains to the firm, but also of windfall losses. Again, bearing in mind the availability of quarterly operating results, and the possibility of staggering the rate increases by class of service, the chances of major windfall gain or loss are reduced to negligible magnitudes. Most important is that the cumbersomeness of rate cases would be reduced as Mr. Goodier hopes, while the spur of regulatory demand for efficiency gains would not be lost.

The specific impact of inflation can be fairly readily estimated directly and aggregated up to an overall impact. While presenting some problems of complexity, this process is not overwhelmingly difficult.

The problems with using productivity and inflation measures, which Mr. Goodier described in his discussion of the CTC inquiry and Order T-474 are real. They are not, however, overwhelming. It is my view that the process of discovering how to streamline the development and application of such measures was prematurely terminated during the unavoidable tumult caused by transition from CTC to CRTC regulation of telecommunications. Such measures can only come close to the "right" answers in terms of rate relief, but they can come so close, while remaining reasonably transparent to third party scrutiny, that they would more than pay for themselves in terms of regulatory costs avoided.

Two further observations deserve to be made in respect of Mr. Goodier's claims. First, in his conclusions to Part A, and to later parts, he asserts that the formula has no disincentive effects. The contrary is true. By removing explicit assessment of productivity the formula has the effect of a cost plus contract with the regulated firm. The disincentive effects of such arrangements are well-known. Second, in Part B, Mr. Goodier asserts that his formula can be applied to particular services. Again he assumes away the problem, by writing up the case as if costs were known. Nearly a decade of CTC/CRTC Cost Inquiry hearings, hundreds of individual service filings, and countless U.S. proceedings make one thing abundantly clear about costs—that is, they are extremely difficult to determine in the simplest of particular cases and impossible to determine by defensible analytic procedures whenever there are significant common costs as there almost always are with individual telecommunications services.

All of this having been said, it remains true that Mr. Goodier has stated the economic tautology with algebraic clarity. That it will not serve the function he proposes for it does not detract from the fact that his statement makes it clear where the approach must be changed to be made operational, and where one might "bite the bullet" and accept solutions to regulatory simplifications in this area which are not perfect. The fact that they are neither perfectly simple nor perfectly accurate does not impair the capacity of those solutions to improve upon the present regulatory situation.

Introductory Comments on the Management Applications of Productivity Analyses

R. E. Olley

The process of measuring productivity gains generates what may be conceived of as a set of economic accounts for a firm. These accounts measure outputs and inputs in terms which are free of price changes. They also measure price changes separately. In doing so, they break inputs down into categorizations which have, at least by intent, some meaningful economic characteristics. These real measures of aspects of a firm's performance over time permit analyses of productivity gain and the causes thereof to be carried out. The results of those analyses provide a picture of gains in the overall efficiency of a firm over time and the variations therein. They also permit those gains to be factored into various components which explain, in some sense, the observed productivity changes.

Now the question arises as to what other uses may be found for what I have called the economic accounts, and what applications may one make of the explanations of productivity gain which are eventually found to be persuasive.

Broadly speaking, there are two types of application which one might expect. Internal to the firm, the economic accounts and explanations of productivity gain may be expected to have some managerial applications. External to the firm, the same bodies of data could be used for regulatory purposes and in the formulation or conditioning of other government policies as they bear on the firm.

The following papers encompass both possible uses of productivity data and analyses. In that sense they carry on directly from the papers presented this morning but in a different direction. The first three (Reimeringer; Chaudry and Burnside; Denny, De Fontenay, and Werner) are concerned primarily with internal managerial uses, which may also have external applications. The fourth paper (Goodier) is concerned with regulatory applications of the economic accounts which are derived.

RESPONSE TO DR. R.E. OLLEY'S COMMENTS ON

THE MANAGEMENT APPLICATION OF PRODUCTIVITY ANALYSES

Ray J. Goodier

Introduction

In his comments on my paper on Rate Adjustment Formula Procedure, Dr. Olley raised issues regarding my position which I will herein attempt to clarify.

Question Addressed

I agree with Dr. Olley that my paper addresses an extremely important question:

> Whether there is any mechanism which can be developed to enable regulators to permit rate increases to regulated utilities when inflationary forces are rampant, making rate applications frequent and rate relief in some sense obviously necessary (~~p. 5~~, l. 3).

Proposed Formula

As Dr. Olley acknowledges, I propose the following formula:

Revenue Adjustment = Test Period Cost, less Base Period
Cost times Real Output Growth Index

This represents the extra revenue that would have made the carrier as well off in the Test Period (i.e., the most recently completed fiscal year) as it had been in the Base Period (i.e., the preceding fiscal year). On the assumption that the carrier will be as badly off in the future as it was in the Test Period, the Commission should permit sufficient rate increases to generate additional annual revenues equal to the Revenue Adjustment.

Application

For example, a rate case in 1981 would use 1980 as the Test Period and 1979 as the Base Period, so that Revenue Adjustment, if applied retroactively, would be just sufficient to have made 1980 as financially successful as 1979. Applied in the future, the Revenue Adjustment would recover the financial position lost between 1979 and 1980.

T-474 Formulation

The formula I suggest is the simplest possible as it requires only the determination of cost and real output growth rate. On the other hand, the T-474 formulation requires the determination of the inflation rate. Considering that this represents extra effort, it makes little sense to favour this formulation. One may argue that the T-474 formulation avoids the need to determine the growth rate. However, since productivity gain requires a knowledge of growth, this line of reasoning would be invalid.

I agree that measurement of cost and growth are perhaps contentious issues, but I fail to see how the introduction of other contentious issues (i.e., rates of inflation and productivity gain) would improve the situation. In any event, cost and growth do not represent insurmountable problems.

Growth Measurments

While real growth is not readily measurable, there are well established procedures that could be refined with time, effort and reasonable cost. With respect to new services, I do not regard these as a problem as the associated costs (measureable according to Phase II of the CRTC Cost Inquiry) represent pure growth costs and would not be permitted as part of the Revenue Adjustment.

Service Costing

With respect to cost measurement, I would agree that as yet the cost of a particular service is contentious (due to the sharing of cost among services), and therefore it is not as yet possible to determine the Revenue Adjustment applicable to individual services. However, I am optimistic that Phase III of the CRTC Cost Inquiry will resolve the problem of Cost Separation By Service (or at least by groups of competitive and monopoly services), and I foresee no reason why my formula could not apply, irrespective of the costing procedures adopted.

Cost Catagories

Meanwhile, my formula could apply, in a global manner, to the categories of cost proposed in T-474, i.e., Operating Wages and Salaries, Taxes (excluding Income Tax), Depreciation, and Other Expenses (i.e. materials, supplies and services used for maintenance and operations, and other expensed items such as rentals, printing, postage, stationery and other general expenses not already provided for in the other three categories).

For example, given the Operating Wages and Salaries for the Base and Test Periods, and the carriers overall rate of output growth, one could calculate the Revenue Adjustment associated with the payroll. Such a Revenue Adjustment would implicity account for payroll inflation offset by labour productivity gains.

Normal accounting and auditing procedures should ensure that the cost categories reported by the carrier are correct. Thus, the above procedure should not be contentious, unless of course, the carrier is deliberately dishonest in misrepresenting the costs incurred, or permitting an unreasonable amount of spare capacity in its operating procedures.

Cost-Plus Contract

Dr. Olley characterizes my proposal as "little more than a blatent cost-plus contract" ([redacted], l. 15). More specifically, Dr. Olley states the following:

> By removing explicit assessment of productivity the formula has the effect of a cost plus contract with the regulated firm. The disincentive effects of such arrangements are well known ([redacted], l. 4).

This is misleading, since any reasonable formula approach to recovering uncontrollable increases in costs, including the T-474 formulation (i.e., Inflation Cost less Productivity Gain), may be described as a cost plus contract. Even the present Rate of Return method of regulation could be so characterized since it implicitly assumes that all costs reported by a carrier may be legitimately passed onto subscribers.

As for disincentive effects, I disagree. The formula would only recover inflation less productivity gain, and therefore would simply maintain the carrier's financial position. In other words, the carrier would be no worse off and certainly no better off than it would have been had inflation and productivity gain not occurred (at least for those costs permitted in the formula procedure).

Incentives for Efficiency

Now, if the regulators wish to encourage carrier efficiency, then another formula could be introduced. As stated in my paper, I propose the adoption of a Productivity Gain formula:

> to provide a basis for determining the additional revenue adjustment required (a fraction of the productivity gain) to sufficiently improve the carrier's financial position to inspire further productivity gains (~~p. 52~~, l. 14).

In my view, even as little as 10% of the Productivity Gain would inspire greater efficiency. This inevitably, would result in a smaller required Revenue Adjustment to maintain the carrier's financial position.

Responsibility of Regulators

Dr. Olley is of the opinion that regulators should test that

> costs have been minimized at current expected grades of service, under current general economic conditions, and with available production techniques (~~p. 7~~, l. 4).

To do this, Dr. Olley suggests that regulators

> establish expected values for overall productivity gain and inflation impact, testing the proposal revenue requirement with these two values(~~p.7~~,1. 14).

As I understand this suggestion, Dr. Olley is proposing the following test formula:

Revenue Adjustment = Prospective Inflation Cost less
Prospective Productivity Gain

This is similar to the T-474 formulation, but is prospective rather than retrospective. As such it implies horrendous difficulties as Dr. Olley himself, has mentioned. Considering that productivity gain requires a knowledge of cost and growth, the formula implies that both cost and rate of growth be forecast, as well as the rate of inflation.

In my opinion, the enormous time, effort and cost (not to mention the lack of precision) associated with this formula would defeat the intent of the Rate Adjustment Formula Procedure in providing a low cost, streamlined approach to price regulation. Perhaps Dr. Olley's proposal would avoid "assuming away the problem" facing regulators, but in my view it would raise more problems than it would solve.

In any event, Dr. Olley's formula would not ensure that costs be minimized. Even with perfect forecasting, Dr. Olley's proposed formula would simply pass anticipated cost increases to the subscribers in the form of increased rates. In the words of Dr. Olley, this would amount to "little more than a blatent (prospective) cost-plus contract". But, of course, Dr. Olley's formula would be no more incentive or disincentive than my own formula approach to price regulation.

Conclusion

Contrary to Dr. Olley's assertions, I maintain that my "RAFP Simple Formula" has no disincentive effects and would provide a low cost, streamlined and reasonably accurate approach to price regulation of Telecommunications.

While not addressing all issues normally covered in a full scale rate hearing, I believe the formula could fulfill the original intention of the Telecommunications Committee of the CTC, i.e.:

> In introducting this proposal (i.e., RAFP), the Committee intends only to decrease the frequency of Rate Hearings. These will still continue to be necessary from time to time (~~T-474, P. 82, L. 23).~~

1981 05 21

INTEGRATION OF FINANCIAL AND ECONOMIC ANALYSIS

THE VALUE OF THE FIRM UNDER REGULATION AND THE THEORY OF THE FIRM UNDER UNCERTAINTY: AN INTEGRATED APPROACH

STYLIANOS PERRAKIS

University of Ottawa

I. Introduction

While the existence of uncertainty in economic decision-making is central to financial theory, the incorporation of such uncertainty in the microeconomic theory of the firm is a comparatively recent phenomenon. A related tendency, which again has only recently been abandoned, was that of separate consideration of production and financial decisions. Consequently, whenever, for instance, financial theory needed elements of the microeconomics of the firm, it tended to adopt the results of production theory under certainty without taking into account the modifications that uncertainty in decision-making was going to bring into these results. This paper examines the interaction of finance and the microeconomics of uncertainty in the particular domain of the theory of the regulated firm, a domain in which both disciplines have traditionally played a major role.

The theory of the firm has paid special attention to regulation ever since the Averch-Johnson (AJ) study alleging input distortions induced by the rate-of-return regulatory constraint. Financial theory, on the other hand, has had a built-in role to play in regulation because of the requirement in the Hope Natural Gas decision that a regulatedfirm's product price provide a "fair" return. In that same decision such a fair return is interpreted as being "commensurate with returns on investment in other entreprises having corresponding risk" and sufficient to "attract capital". Hence, financial theory must provide rules determining the allowed percentage rate of return and the rate base as well as evaluate the random income streams that the regulated firm generates in order to determine the impact of regulation upon the value of the firm.[1] These random income streams are determined by the production decisions of the firm. At the same time, these production decisions take place under a regulatory constraint, whose determination by the regulator needs inputs from the capital markets' evaluation of the income streams of the regulated firm. Hence, financial and production theory have to be examined in conjunction during the study of the behavior of a regulated firm.

Surprinsingly, this did not happen in most studies[2], and for the most part the microeconomic discussions of the problems of the regulated firm took place in the absence of financial considerations, and vice-versa. The impact of such a separation was, in our opinion, more serious on the financial side, since most studies in the financial literature seem to have ignored the refinements that have appeared in the theory of the regulated firm. In particular, major controversies concerning the nature of the regulatory process itself (which will be shown to have an impact upon the value of the regulated firm) have been bypassed in the financial literature. Similarly, even within the narrow confines of the view of regulation adopted by most authors there are a number of questionable assumptions and derived results that need to be reexamined. Thus, the valuation of the uncertain income streams generated by a firm subject to rate-of-return regulation will be the main focus of this paper.

On the production side the development in recent years of the microeconomics of uncertainty has forced the consideration of financial decisions, given that the objectives of the corporate firm under uncertainty are not easily defined without taking into account the financial side of the firm. Unfortunately, for reasons that will be examined further on, most existing models of simultaneous production and financial decision-making are not easily adaptable to regulated firms. In addition, the one-period static equilibrium nature of these models does not lend itself easily to the analysis of "real-life" uncertainty cases, which are characterized by multiple information lags and intertemporal decision-making. Consequently, the problem of the proper objectives of the firm under uncertainty will not be examined here in detail but it will be assumed according to existing models that the firm maximizes value, expected profit, or expected utility of profit.

In the financial literature many important contributions came as a result of the Miller-Modigliani (MM) empirical study [28] of the cost of capital in the electric utility industry. Although this study was an attempt to apply the earlier MM theory ([29] [30]) and was not directly concerned with regulation, the fact that the theory was applied to a regulated industry provided several discussions and controversy ([8], [9], [13], [14], [16], [17], [19]) for a decade. This controversy was extremely fruitful, because it helped bring into the foreground the microeconomic model that determines the earnings stream of the regulated firm. The formulation of this model was due to Gordon and his associates ([9], [16], [18]), and it was basically a reinterpretation of the AJ model of regulation under certainty, with expectations replacing the deterministic streams of profit in the objective function and the constraint. The microeconomic consequences of this reinterpretation were examined recently by Meyer [27], and they were shown to parallel fairly closely the deterministic AJ results. Nonetheless, even within the context of the AJ-Gordon model the financial questions are far from being completely solved. The principal problem is the nature of the uncertainty in the earnings stream and the resulting consequences upon the value of the firm. These consequences are, in turn, dependent upon the model of valuation used. The main models examined were the Capital Asset Pricing Model (CAPM [2], [19], [27]), the simultaneous production and financial model based on the spanning property of the earnings stream ([3], [22]) and the MM cost of capital model ([8], [13], [14], [17], [19]).

The fundamental characteristic of the AJ-Gordon model of regulation is that it is "forward-looking" in essence, given that what is constrained within that model is the (by definition unobservable) mathematical expectation of earnings. Hence, it requires prefectly shared information and expectations between firm and regulatory agency, as well as a test rule that is not based on observed past performance. Otherwise, if regulation of future performance is based on observed firm behavior during a past test period, the firm has an incentive to tailor its performance to fit regulatory expectations. Such "backward-looking" regulatory models have appeared quite frequently in the microeconomic literature of uncertainty.[3] Their justification can be found in detailed studies of the regulatory process such as those of Joskow ([20], [21]). The implications of backward-looking regulation for the value of the firm have not been explored until now; they form the main result of this paper.

In the next two sections a one-period model of the firm under forward-looking regulation is presented, and some of the earlier valuation results are re-examined in the context of new developments in the theory of the firm under uncertainty. These developments in turn, raise questions about the robustness and generality of some of the valuation results, although all valuation models remain valid under special conditions. It is also pointed out that the functional forms of the earnings function depend upon the length of the regulatory lag relative to that of the static period considered.

In section IV backward-looking regulation is examined, given that a forward-looking valuation model of the firm has been established. By using Ross' arbitrage-based approach to the valuation of risky streams ([40], [41], [42]), it is shown that the value of the regulated firm under various forms of backward-looking regulation can be derived by combining financial instruments of this same firm under forward-looking regulation, in combination with simple options (calls or puts) on these instruments. A number of old results are shown to hold for the backward-looking regulated firm, and some new results are also presented.

In the final sections of this paper it is attempted to extend the model of backward-looking regulation beyond the single-period framework. Some results are derived under simplified assumptions. It is pointed out, however, that the main problems lie in the area of the multiperiod valuation of uncertain income streams. Hence, they are common to forward-looking regulation as well, and the relation between the values of firms under forward- and backward-looking regulation is not dependent on as yet unsolved aspects of multiperiod valuation.

II The General Model

We denote by p the product price and let Q(p,u) denote the random demand curve, where u is a random factor. The firm's capital assets is denoted by K, s is the allowed rate of return and $r < s$ is the riskless rate of interest. t is the tax rate and D the amount of debt in the case of levered firm. Subscript L and no subscript denote levered and unlevered firms respectively.

The firm selects simultaneously the size of its assets K, the output price p, and the other production inputs by maximizing profits while keeping the rate of return on assets at or below s. The cost of capital r, at which K is valued during profit maximization, is exogenous, and less than s. Simialrly, the regulatory rule is such that the profit stream is different for levered than for unlevered firms. Hence, the choices of the regulated firm are going to depend on leverage.

Let x_i, w_i, i = 1,... n denote the variable production inputs and their prices for the regulated firm. The non-levered firm under certainty chooses its inputs and its output price by solving the following problem

$$(1) \qquad \text{Max } \{(1-t)\ [pQ(p,u) - rK - \sum_{i=1}^{n} w_i x_i]\}$$

subject to the production function and rate-of-return constraints $Q \leq F(x_1,.., x_n,K)$ and $(1-t)[pQ(p,u) - \sum_{i=1}^{n} w_i x_i] \leq sK$ respectively. It is well-known that the firm minimizes the cost of the variable inputs given K, but that K is chosen in a non-total cost minimizing way. Hence, if C(Q,W,K) denotes the variable cost function, where the vector $W \equiv [w_1,.., w_n]$, the problem becomes

$$(2) \qquad \underset{K,p}{\text{Max}}\ \{(1-t)\ [pQ(p,u) - rK - C(Q(p,u), W,K)]\}$$

subject to

$$(3) \qquad (1-t)\ [pQ(p-u) - C(Q(p,u), W,K)] \leq sK$$

The solution is easier to visualize if we define

$$(4) \qquad N(u, W, K) \equiv \underset{p}{\text{Max}}\ \{pQ(p,u) - C(Q(p,u), W, K)\}$$

the quasi-rents function or the firm. It can be shown that under certain common assumptions about the firm's revenue and production functions the function N is concave and increasing in K. Hence, under certainty and in the solution of (2) the constraint (3) is satisfied with equality, and the optimal choice of K is the positive solution[4] of the equation in K $N(u,W,K)=\frac{sK}{1-t}$. We note that this solution is independent of r as long as r < s.

For the levered firm the maximand in (1) is augmented by the term trD, where the amount of debt D is assumed exogenous.[5] This does not affect the analysis until the last step, in which the total assets K_L are determined by solving the equation in K $(1-t)N(u,W,K) + trD = sK$. With the assumed shape of N it follows automatically that $K_L > K$. The comparison of p_L and p is not straightforward and depends to some extent on the properties of the cost function $C(Q,W,K)$, i.e. on the production function $F(x_1,.,x_n, K)$. A commonly accepted assumption (especially in the case of electric utilities) is the so-called *putty-clay-hypothesis*[6], according to which, although input substitution may be feasible *ex ante* (before K is selected), no such substitution is allowed *ex post* and the production function becomes of the fixed coefficients type. This means that $C(Q,W,K) = [\sum_{i=1}^{n} w_i b_i \ (K)] \ h(Q)$, where we assumed that the production structure is homothetic, i.e. the cost function is separable into a product of two functions, of which one contains only Q. In financial analysis it is also often assumed[7] that the production function exhibits constant returns to scale in the variable inputs, i.e. that $F(x_1,..,x_n,K)$ is linear homogeneous in $(x_1,.., x_n)$. This is strictly equivalent to $h(Q)=Q$, i.e. that average variable cost is constant. With such an assumption it is fairly easy to show that $p > p_L$, based on the fact that $\sum_{i=1}^{n} w_i b_i (K_L) < \sum_{i=1}^{n} w_i b_i \ (K)$, given that $K_L > K$ and $C(Q,W,K)$ is decreasing in K. Otherwise the relative size of p and p_L depends on the second derivatives $\frac{\partial^2 C(Q,W,K)}{\partial Q \partial K}$, $\frac{\partial^2 C(Q,W,K)}{\partial Q^2}$ and $\frac{\partial [pQ(p,u)]}{\partial p^2}$.

III The Value of the Firm Under Forward-looking Regulation

In order to formulate the model under uncertainty we assume that there is a single time interval, at the end of which (time 1) the firm is disolved. All uncertainty is revealed at the end of the period, at which point the variable inputs x_i, i=1,.., n are selected and the output produced instantaneously. The capital stock K is selected at the beginning of the interval (time 0) and we assume that its price is unity. Uncertainty appears in the random factor u of the output demand curve, as well as in the variable input price vector W, which is unknown at the time K is selected. As before, the amount of debt D and the rate of return s are exogenously determined.

With this specification it can be shown that in all valuation models the first step of the certainty analysis (leading from (1) to (2)) remains unchanged. Similarly, forward-looking regulation in this context implies that the regulatory constraint (3) is satisfied with equality only when the expectation is taken in the LHS. This implies perfect agreement and sharing of information between firm and regulator on the future states of the world, as well as a point estimate of the upper limit s of the regulatory constraint.

A number of alternative objectives of the regulated firm in this uncertain world have appeared in the literature, some of them explicitly formulated in a simultaneous production and financial equilibrium, and others implicit in the context of different problems. They will be surveyed briefly within the framework of the basic problem of finding the value of the firm.

a) Expected profit maximization: These are not, strictly speaking, models of the firm that lead to valuation, but they form the obvious extension of the theory of the firm under certainty. As such, they have been used extensively (though implicitly), and almost exclusively in the MM cost-of-capital for regulated firms controversy ([8], [13], [14], [17], [19]). There are two limiting versions of this model depending on the institutionally determined regulatory lag. If the lag is "small" relative to the length of the static period then the output price selection takes place after uncertainty has been resolved, while capital K is selected initially under uncertain earnings by satisfying the expected rate of return constraint. If the lag is "large" then output price is selected by expected profit maximization simultaneously with the size of the capital stock.[8] Both versions yield ultimately similar conclusions with respect to the cost of capital, although the simultaneous price-capital determination complicates the analytical formulation.

The cost-of-capital controversy as expressed especially in [13], [14], [16], [17], and [19], refers to the validity for regulated firms of the well-known MM formula $\rho_L = \rho + (1-t)\ (\rho-r)\ \frac{D}{S}$, where ρ and ρ_L represent the cost-of-capital without and with leverage, and S is the value of the levered firm's equity. It was asserted by Gordon in [16] and disputed by Elton and Gruber (EG), in [13] and [14] that the MM formula does not hold in regulated industries under most common specifications of uncertainty. In [19] the arguments were re-examined and it was concluded that the MM-EG arguments under regulation are valid under special (but commonly assumed) circumstances, such as an output demand curve with multiplicative uncertainty; otherwise the relation between ρ_L and ρ is a nonlinear function dependent on leverage. However, under the more general formulation followed here the MM formula does not hold even in the cases examined in [14] and [19]. This occurs because of two features ignored in these previous studies: the random nature of the firm's cost function and the fact that K_L is > K, combined with the nonlineraity of earnings with respect to the capital stock.[9]

To demonstrate this we denote by V and $V_L \equiv S+D$ the values of the unlevered and levered firms respectively, and we adopt the notation of [42] by denotingby < > the value of the random cash flow in the brackets. It was shown in [42] that within the context of the theory of arbitrage valuation ([40], [41]) the valuation operator < > is linear, and that the MM theory is a special case of arbitrage valuation. With this notation we obviously have

(5a) $V = (1-t) < N(u,W,K) >$

(5b) $V_L = (1-t) < N(u,W.K_L)> + tD$,

when price is determined ex post, or

(6a) $V = (1-t) [< pQ(p,u) > - < C(Q,W,K) >]$,

(6b) $V_L = (1-t) [< p_L Q(p_L,u)> - < C(Q_L,W,K_L) >] + tD$,

for price determined simultaneously with capital stock, and for $Q_L \equiv Q(p_L,u)$. In the first case K and K_L are determined from $(1-t)E[N] \equiv (1-t)\bar{N} = sK$ or $(1-t)E[N_L] \equiv (1-t)\bar{N}_L = sK_L - trD$, while in the second case expected profit maximization determines p and p_L by the equality of marginal revenue and marginal cost. In both cases it will be shown by a simple (but realistic) counter-example with a constant elasticity demand curve and multiplicative uncertainty that the MM formula does not hold.

From the definitions of ρ_L and ρ it follows always that $\rho_L = \frac{\bar{N}_L}{N} \rho \frac{V}{S} - (1-t)r\frac{D}{S}$. Hence, the MM relation $\rho_L = \rho + (1-t)(\rho-r)\frac{D}{S}$ holds iff $\frac{\bar{N}_L}{\bar{N}} \frac{V}{S} = 1 + (1-t)\frac{D}{S}$, or $\frac{\bar{N}_L}{\bar{N}} \frac{V}{S} = \frac{1}{S} <(1-t)N_L >$ or, since $V = <(1-t)N>$, iff $< \frac{N}{\bar{N}} > = < \frac{N_L}{\bar{N}_L} >$ ([19], p. 707). However, it is very easy to find counter-examples ivolating this equality. Suppose, for instance, that $Q = Bup^{-2}$, and that $C(Q,W,K) = Q[w_1 b_1(K) + w_2 b_2(K)]$, with w_2 fixed and w_1 random.

Then $N(u, W,K) = \frac{Bu}{4} [w_1 b_1(K) + w_2 b_2(K)]^{-1}$. The MM relation holds iff

$$\frac{< u[w_1 + w_2 \frac{b_2(K)}{b_1(K)}]^{-1} >}{E[u[w_1 + w_2 \frac{b_2(K)}{b_1(K)}]^{-1}]} = \frac{< u[w_1 + w_2 \frac{b_2(K_L)}{b_1(K_L)}]^{-1} >}{E[u[w_1 + w_2 \frac{b_2(K_L)}{b_1(K_L)}]^{-1}]} .$$

In general $\frac{b_2(K)}{b_1(K)}$ is not constant with respect to K. For instance, if inputs 1 and 2 are fuel and labor respectively it may be expected that the ex ante substitutability of capital is greater for fuel than for labor, implying that $\frac{b_2(K_L)}{b_1(K_L)} < \frac{b_2(K)}{b_1(K)}$ if $K_L > K$. Similarly, when p and K are determined simultaneously and $\overline{uw_1} \equiv E(uw_1)$the expected profit maximizing price for the unlevered firm is $p = 2[\overline{uw_1}b_1(K) + w_2b_2(K)]$. As before, we must compare the value of the before tax earnings normalized by their expected value for levered and unlevered firms. This value for the unlevered firm is equal to $\frac{< Bup^{-2} - Bup^{-2}(w_1b_1(K) + w_2b_2(K)) >}{Bp^{-1} - Bp^{-2}(\overline{uw_1}b_1(K) + w_2b_2(K))}$, which, by the linearity of the valuation operator and for the expected profit-maximizing price is equal to $< u > - \frac{< u(w_1b_1(K) + w_2b_2(K)) >}{\overline{uw_1}b_1(K) + w_2b_2(K)}$. Hence, the MM relation holds

iff $\frac{< u(w_1 + w_2 \frac{b_2(K)}{b_1(K)})>}{\overline{uw_1} + w_2 \frac{b_2(K)}{b_1(K)}} = \frac{< u(w_1 + w_2 \frac{b_2(K_L)}{b_1(K_L)}) >}{\overline{uw_1} + w_2 \frac{b_2(K_L)}{b_1(K_L)}}$, which does not hold in general as explained above since K_L is > K and the substitutability of the ex post inputs for capital is not the same.

The conclusion, therefore, is that the validity of the MM cost-of-capital theory for regulated industries is extremely restricted, even under the commonly adopted econometric assumptions such as those of multiplicative demand uncertainty as in [19]. The randomnes in input prices makes the average variable cost uncertain,[10] while regulation, which induces the choice of a larger capital stock for a levered firm, brings systematic differences in the probability distributions of the streams of earnings between levered and unlevered firms. Thus, Gordon's objections to the use of the MM theory for regulated firms are valid, even though the alternative formula that he proposed is also flawed, as shown in [13] and [19].

A common assumption used for the justification of the MM theory is that N and N_L are perfectly correlated ([28], note 3, [19], p. 708). A sufficient condition for this perfect correlation when both u and W are random is the separability of N(u,W,L) into the form n(u,W) g(K), which would obviously validate the MM cost-of-capital theory. This separability, unfortunately, implies rather restrictive conditions[11] on the shape of the demand and production functions of the firm, and it also plays a role in other valuation theories to be examined below.

b) Value maximization in a mean-variance world: Let R_m denote the return to the market portfolio, and λ the market price of risk. Under the assumption of *ex post* determination of output price the analysis under the CAPM is very similar to the expected profit case: the capital stock is determined by the rate-of-return constraint. The before-tax value of the unlevered firm then becomes

$$(7) \qquad < N(u,W,K) > \equiv \frac{\bar{N} - \lambda \mathrm{Cov}(N,\ R_m)}{1 + r}$$

and V = (1-t) < N > for this unlevered firm. The effect of leverage is identical to that of the previous case, the key element in the comparaison of the distributions of $\frac{N}{\bar{N}}$ and $\frac{N_L}{\bar{N}_L}$ being their covariance with the market index return R_m. These covariances are, in general, unequal, as it can be easily seen in the counterexamples presented above.[12]

If p and K are determined simultaniously then the analysis becomes more complex, and additional assumptions are needed in order to preserve the certainty AJ model. We denote by $R \equiv R(p,u) \equiv pQ(p,u)$ and $C \equiv C(Q,W,K)$ the revenue and cost functions respectively, and let a bar represent the expectation. Then the before tax value of the all equity firm becomes

$$(8) \quad \langle R-C \rangle \equiv \frac{\bar{R} - \bar{C} - \lambda \, \mathrm{Cov}(R-C, R_m)}{1 + r} ,$$

with $V = (1-t) < R-C >$, as before. Value maximization introduces a number of difficulties, some of which have already been discussed elsewhere in a different context, while others are peculiar to rate of return regulation. For instance, the choice of K in a general equilibrium mean-variance framework implies that the market price of risk parameter λ is affected in a predictable manner by the firm's decisions. This resulting complications are generally avoided by adopting the *competitivity* assumption[13], under which λ is assumed constant to reach the Pareto-optimal solution, If this assumption is adopted then we have the following first-order conditions for the regulated firm.

$$(9a) \quad \frac{\partial V}{\partial p} + \mu(1-t)(- \frac{\partial \bar{R}}{\partial p} + \frac{\partial \bar{C}}{\partial p}) = 0$$

$$(9b) \quad \frac{\partial V}{\partial K} - 1 + \mu \, [s + (1-t) \frac{\partial \bar{C}}{\partial K}] = 0$$

$$(9c) \quad \mu \, [s \, K-(1-t)(\bar{R} - \bar{C})] = 0$$

where $\mu \geq 0$ is the Kuhn-Tucker multiplier associated with the rate-of-return constraint and the price of capital was normalized and set equal to 1. The unregulated firm's choices are given by (9a,b) for $\mu = 0$.

If the rate-of-return constraint is effective then μ must be > 0 at the regulated (p,K). Assume that $\frac{\partial^2 V}{\partial K^2}$ and $\frac{\partial^2 V}{\partial p^2}$ are < 0, and that $\frac{\partial^2 V}{\partial p \partial K} > 0$ at a neighborhood of the unregulated solution and that the regulated firm's choices fall in that neighborhood. If the firm's production function is quasi concave in all the inputs then the variable cost function is convex in K for all (Q,W). This means that the equation $sK = (1-t)[\bar{R} - \bar{C}]$ has two solutions if regulation is effective, of which the larger will be the one used by the firm.[14] Hence, s will be $> -(1-t)\frac{\partial \bar{C}}{\partial K}$. It follows from (9a, bc) and the assumed signs of the second derivative that the AJ results under certainty, namely that under rate-of-return regulation the output price will be lower and the capital stock higher than in the absence of regulation, are also valid for this type of firms. As a corollary, it is noted that the conclusions of the expected profit-maximization model with respect to leverage and the cost of capital are unchanged in a mean-variance framework.

This value-maximizing model of the regulated firm within the CAPM is the one, for which the extension of the AJ results in the domain of uncertainty maintains most of the certainty conclusions. In addition, and in spite of its weak theoretical foundations, it is the best known and most widely accepted of all financial models. It is, therefore, surprising that even for this model the robustness of the conclusions disappears when the regulatory specifications are varied.

c) Other simultaneous production and financial equilibrium models: such models are based on extensions of the Arrow-Debreu general equilibrium models to an economy with incomplete markets for state-contingent claims. These extensions take place by restricting the shape of the earnings function of the firm. They will be treated very briefly, since the resulting valuation models (other than the CAPM) are very restrictive in the case of regulated firms and have had very little impact on applied financial research.[15]

We consider for simplicity the case of insignificant regulatory lag, in which the value of the unlevered firm is equal to (1-t) < N(u,W,K) >. An essential condition for the existence of a unanimously preferred simultaneous production equilibrium is that the earnings function N(u,W,K) satisfy the so-called <u>spanning condition</u>([2], [12], [23]):let J be the number of risky firms in the economy, and the subcript j denote the monopolist. Then the spanning condition implies that there exists a set of J coefficients a^j_h for the j^{th} firm independent of (u,W) (but not necessarily of the capital stocks of the J firms) such that[16]

$$(10) \qquad \frac{\partial N_j(u,W,K_j)}{\partial K_j} = \sum_{h=1}^{j} a^j_h \, N_h(u,W,K_h) + H_j(K_j) \; ,$$

where $H_j(K_j)$ is a known function, constant in (u,W).

If this condition holds then the value of the firm becomes a function of the entire matrix of coefficients (a^m_h), m , h = 1,..,J. To prove this it is easiest to adopt the discrete state-space formulation of [12]. Let k = 1,.., K denote the set of possible values of (u, W) and Ω the J x J matrix (a^m_h) of known coefficients. Following [2] and [36] we also denote by ω^i the K-vector of normalized marginal utilities of the i^{th} consumer weighed by the probabilities of occurrence of each state, evaluated at the joint production - financial equilibrium. Then, the following relationships hold in our notation, as in [12], for any $m = 1,., J$

(11a) $(1-t)[\nabla N_{mk}]\ \omega^{i} = [(1+r)]$

(11b) $[V_{m}](1+r) = (1-t)[N_{mk}]\ \omega^{i}$

(11c) $[\nabla N_{mk}] \ = \Omega\ [N_{mk}] + H[1]$

where H is a J x J diagonal matrix with $H_{m}(K_{m})$ its m^{th} element, $\nabla N_{mk} \equiv (\frac{\partial N_{mk}}{\partial K_{m}})$, the brackets indicate an appropriately dimensioned matrix of the elements within the brackets, and (11c) is a matrix version of (10). Then, since $\sum_{k}\omega^{i}_{k} = 1$, we have $[1+r]=(1-t)\Omega[N_{jk}]\omega^{i} + H[1](1-t)$ from which we get, because of (11b):

(12) $$V_{j} = \Omega^{-1}[1] - \frac{(1-t)}{1+r}\,\Omega^{-1}H[1]$$

where H[1] is obviously the vector with $H_{m}(K_{m})$ as its m^{th} element.

In practice the spanning condition (10) has been applied under the following more restrictive version

(13) $$N_{j}(u,W,K_{j}) = n_{j}(u,W)g_{j}(K_{j}) + h_{j}(K_{j}),$$

which can be shown to imply (10) for the following values of Ω and H: $a^{j}_{h} = 0$, $h \neq j$, $a^{j}_{j} = \frac{g'_{j}(K_{j})}{g_{j}(K_{j})}$, $H_{j}(K_{j}) = h'_{j}(K_{j}) - \frac{g'_{j}(K_{j})}{g_{j}(K_{j})}\ h_{j}(K_{j})$. Then the value V_{j} of the firm is equal to $\frac{g_{j}(K_{j})}{g'_{j}(K_{j})} - \frac{(1-t)}{1+r}[\ \frac{h'_{j}(K_{j})g_{j}(K_{j})}{g'_{j}(K_{j})} - h_{j}(K_{j})]$. The analysis is almost identical when p and K are determined simultaneously, with the difference that the *ex ante* decision variable K_{j} becomes now a two-dimensional vector (p_{j}, K_{j}).

Equation (12) establishes the value of the firm in the absence of a regulatory constraint. As pointed out in [2] (p. 214), with free entry and perfect mobility of resources the RHS of (12) will be equal to K_j and the firm will not be able to generate rents for its owners. The key to this model and to all its variants ([2], [3], [12], [22], [23]), is the fact that the vector ω^i is unaffected by a change in the capital stock. K_j, implying that investors act as price-takers with respect to their implicit valuation of a dollar of return in a given state.

The analysis of such a model under a rate-of-return constraint has been examined in detail in [3] and will not be repeated here. It is sufficient to state that under "naive" rate-of-return regulation (the types used here, meaning that output pricing decisions are supposed to conform in the firm's perceptions to an expected profit regulatory constraint) the AJ certainty results hold. It is also pointed out that if by contrast regulation takes place in a "sophisticated" manner (i.e that the firm does not realize the link between pricing decisions and its own choice of capital) then no technical inefficiency need result. Unfortunately, the practical difference between naive and sophisticated regulation is unclear.

The most serious drawback of these models is that the separability condition (13), or its equivalent when $pQ(p,u)-C(Q,W,K)$ replaces $N(u,W,K)$, are not satisfied in general with most commonly used production and demand specifications when both u and W are random.[17] Consequently, and barring very restrictive assumptions, these capital market theories do not provide as yet valuation models for regulated firms.

As a general conclusion for the forward-looking regulatory models we note that they all share a number of characteristics: they are single-period models, whose multiperiod extension presents a number of difficulties. They do not preserve the MM cost-of-capital results except under very restrictive separability conditions. Finally, they do preserve most of the AJ certainty results under "acceptable" assumptions. In the next section it will be shown that under backward-looking regulation this last property disappears even though precise relationsmay be established between the values of the firm under backward and forward-looking regulation.

IV Backward-looking Regulation and the Value of the Firm

This type of regulation was described in detail in articles by Joskow ([20], [27]), which contained extensive criticisms of the traditional AJ model. Within the context of this paper the fundamental difference between the Joskow and the AJ - Gordon views of the regulatory process is that regulation in the former is based on observed past performance, rather than expected future performance as in the latter. The regulated firm, knowing this, adjusts its own performance in anticipation of the regulatory action.

In the Joskow model there are two regulatory constraints, a lower, as well as an upper, rate of return limit. Firms whose realized earnings approach the lower limit know that they have good chances of getting a rate increase. Firms whose earnings are in excess of the upper limit run the risk of a regulatory hearing initiated by the regulatory commission under the prodding of consumer interests. This hearing, in addition to being costly and time-consuming, will force the firm to earn below the upper limit, probably on the lower rate-of-return boundary. Between the two limits the firm will operate under a stable product price without any rate-of-return consideration ([20], pp. 133-134).

In such a world it would make sense for the firm to try to voluntarily limit its earnings in order not to exceed the upper limit. In Joskow's model the firm does this by voluntarily decreasing its output prices. An equally plausible (and easier) way of reducing earnings is by operating above the minimum cost level. Although the efficiency implications of these two methods are very different, the results from the point of view of observed earnings and value of the regulated firm are similar.

Two special cases of the general Joskow model have also appeared in the literature of the regulated firm under uncertainty. The first ([33], [34], [35]) assumes only an upper limit and the second [10] only a lower limit on the realized rate of return. Both can be treated within the context of the model developed in this section.[18] On the other hand the general Joskow model of backward regulation is applicable only when p and K are simultaneously determined in the forward case. This because when the output price is chosen in a profit-maximizing way and the earnings are equal to N(u,W,K) there are no instruments available to raise these earnings to the level defined by the lower limit on the rate of return. Hence, we shall assume for the general case that the before-tax earnings under forward-looking regulation are equal to pQ-C(Q,W,K), where p and K are fixed.

Let $V_f(p,k) \equiv (1-t) < pQ-C(Q,W,K) >$, where the valuation operator < > may not necessarily be unique if mean-variance or spanning are not assumed; otherwise, the appropriate model of the previous section may be used. Likewise, denote by V_b the corresponding value of the backward-looking regulated firm with the same capital stock, initial output price, demand and technology. Instead of the regulatory constraint (3) we now have the following mechanism. Let s_1 and $s_2 > s_1$ denote the two limits on the realized rate of return. If the firm's earnings at minimum cost fall below s_1K then the output price is automatically adjusted to bring them to s_1K. If, on the other hand, these earnings are above s_2K then the firm reduces output price or does "cost-padding" in order to bring them to s_2K and avoid a rate hearing that may bring the earnings further down to s_1K. Otherwise the price remains unchanged and production occurs at minimum cost. All adjustments are assumed "frictionless" (no lags).

For a Joskow-type regulation this firm behavior is clearly optimal for the firm if the probability of a rate hearing when the earnings exceed s_2K is equal to 1. Alternatively, it can be shown that a self-imposed upper limit s_2 exists if this probability is an increasing function of the difference between observed earnings and s_1K. This self-imposed upper limit implies an (undoubtedly correct) assumption that the firm acquires information about its own realized costs and revenues before the regulators.

We shall solve the valuation problem in detail for a single period first, and then we shall extend it recursively to more than one periods. The extension is difficult, principally because of the sequential pricing mechanism.[19] A "period" in our context is defined as a time interval, during which at most one price change can take place. Hence, our single-period formulation is also valid for "long" time intervals of product price stability, such as the late fifties and early sixties.

The random variable $\tilde{X} \equiv [p\ Q - C(Q,W,K)]\ (1-t)$ takes values in an interval on the real line, the distribution being assumed discrete (for simplicity and without loss of generality), i.e $\tilde{X}\ \varepsilon[X_o, X_o + m\Delta X]$, where $X_o \equiv \inf_{(u,W)} [p,Q - C(Q,W,K)](1-t)$, and ΔX is the step size of the discrete distribution. Under the assumed regulatory mechanism the value V_b may be derived from $V_f(p,K)$ by using the theory of contingent claims prices.

We define the integers k_1 and k_2 so that $X_o + \Delta X = s_i K$, $i = 1, 2$, $k_1 < k_2$. The random returns accruing to the stockholders of the regulated firm are equal to s_1K for $\tilde{X}\ \varepsilon[X_o, X_o + (k_1-1)\Delta X]$, to $\tilde{X}$ for $\tilde{X}\ \varepsilon[X_o + k_1 \Delta X, X_o + k_2 \Delta X]$, and to s_2K for $\tilde{X}\ \varepsilon[X_o + (k_2 + 1)\Delta X, X_o + m\Delta X]$. This pattern of returns may be replicated by means of the following portfolio: a risk-free investment yielding s_1K after one period, a long position in European call options on the value $V_f(p,K)$, with an exercise price of s_1K, and a short position in European call options on $V_f(p,K)$ but with exercise price s_2K. The first call option pays 0 when $\tilde{X} \leq s_1K$, and $\tilde{X} - s_1K$ otherwise. The second call option pays 0 when $\tilde{X}\ \varepsilon[X_o, X_o + k_2\Delta X]$ and $\tilde{X} - s_2K$ otherwise, hence the portfolio replicates exactly the returns of the regulated all-equity firm. By the arbitrage theory of valuing risky streams the value of the backward-looking regulated firm is, therefore, equal to the value of the portfolio. If $O(S,Y,n)$ denotes the value of European call option on equity whose value is S, exercise price Y and expires in n periods then we have

$$(14)\quad V_b = \langle s_1K\rangle + O(V_f', s_1K,1) - O(V_f', s_2K,1)$$

$$= \frac{s_1K}{1+r} + O(V_f', s_1K,1) - O(V_f', s_2K,1)$$

where $V_f' \equiv V_f'(p,K)$ and r is the one-period riskless rate of interest.

Equation (14) is a fundamental result of this paper that yields the value of a "backward-looking" regulated firm with a given price and assets as a function of the value of the forward-looking firm with the same price and assets. We note that this equation is valid both before and after taxes, with suitable reinterpretation of the symbols. Since the valuation operator in the arbitrage theory is linear, and the call price O(S,Y,n) is linear homogeneous in (S,Y), it suffices to reinterpret V_f, s_1 and s_2 as before-tax value and limits on rate-of-return respectively.

An implied assumption of arbitrage theory that is necessary for the uniqueness of the valuation operator < > is that both backward and forward-looking regulated firms, as well as the call options on V_f, coexist in the same capital market ([42], p. 459). This is, of course, not true, since only one of the two views of regulation is true at any particular time. However, equation (14) is still true for any valuation operator in the following sense: for any given V_f, if an identical backward-looking regulated firm were to be established in the same market and a capital market equilibrium (in the lack of arbitrage opportunities sense) were to be established, then (14) would hold.

Suppose now that the lower limit s_1K disapears and that only the upper limit s_2K exists as in [33], [34] and [35]. Then the earnings of this firm are now $\tilde{X}$ for $\tilde{X} \varepsilon [X_o, X_o + k_2 \Delta X]$, s_2K otherwise. Hence, the earnings stream can be duplicated by a portfolio containing V_f and short calls on V_f with exercise price s_2K. It follows then that the value $\hat{V}_b$ of the firm is

$$(15) \quad \hat{V}_b = V_f - O(V_f, s_2K, 1)$$

With the same reasoning, if only the lower found s_1K is kept as in [10] the value of the corresponding firm $\bar{V}_b$ is

$$(16) \quad \bar{V}_b = \frac{s_1K}{1+r} + O(V_f, s_1K, 1)$$

The well-known inequality ([25], p. 144) $O(S,Y,1) \geq \text{Max}\ \{0, S - \frac{Y}{1+r}\}$, when applied to (14) - (16) establishes clearly that $\bar{V}_b \geq V_b \geq \hat{V}_b$.

Equations(14)-(16), being based on contingent claims pricing theory, are consistent with several asset pricing and option pricing models. Yet, in spite of their generality they may be used for the derivation of a number of results under appropriate assumptions. Suppose, for instance, that we wish to ascertain whether the AJ certainty results hold for backward-looking regulated firms given that they hold for forward-looking firms. Assume as before that $\frac{\partial^2 V_f}{\partial K2} < 0$ at the chosen or regulated price, and that $\frac{\partial^2 V_b}{\partial K^2} < 0$ at a neighborhood of the optimal solution. Assume also that we are in a value-maximizing world. Since in defining $V_f \equiv (1-t) < pQ - C(Q,W,K) > \equiv <\tilde{X}>$ for fixed p and K we did not assume any mechanism for determining the values of these parameters, we may now assume that p and K have both been determined by unconstrained maximization of $V_f - K$. Then, if p is kept constant at that level but K is varied in order to satisfy an expected profit rate-of-return constraint $E[\tilde{X}] \leq s_2K$ that is binding at the unconstrained optimum, it is easy to see that we get a capital stock for the forward-looking regulated firm that is larger than that derived by unconstrained value maximization. However, this excess capitalization does not hold for the backward-looking regulated firm that maximizes $V_b - K$ at the same output price. To see this we compute the quantities $\frac{\partial V_b}{\partial K} - 1$ or $\frac{\partial \hat{V}_b}{\partial K} - 1$ at the value of K, at which $\frac{\partial V_f}{\partial K} - 1 = 0$. From (14) and (15) we have

(17a) $$\frac{\partial V_b}{\partial K} = \frac{s_1}{1+r} + \left[\frac{\partial O_1}{\partial V_f} - \frac{\partial O_2}{\partial V_f}\right] \frac{\partial V_f}{\partial K} + s_1 \frac{\partial O_1}{\partial (s_1 K)} - s_2 \frac{\partial O_2}{\partial (s_2 K)}$$

(17b) $$\frac{\partial \hat{V}_b}{\partial K} = \frac{\partial V_f}{\partial K}\left[1 - \frac{\partial O_2}{\partial V_f}\right] - s_2 \frac{\partial O_2}{\partial (s_2 K)}$$

where $O_i \equiv O(V_f, s_i K, 1)$, $i = 1,2$.

For $\frac{\partial V_f}{\partial K} = 1$ (17a,b) may be $\gtrless 1$ depending on a number of factors, including the distributional characteristics of $\tilde{X}$. As a specific example, we consider (17b) when the function $O(S,Y,n)$ is given by the well-known Black-Scholes [6] option pricing model. Then, if r is redefined to indicate the instantaneous rate of interest in continuous time, N_m ( ) is the cumulative standard normal distribution, and σ is the standard deviation of $\tilde{X}$, we have ([43], p. 24).

(18) $$\frac{\partial \hat{V}_f}{\partial K} - 1 = s_2 e^{-r} N_m\left[\frac{\ln V_f - \ln(s_2 K) + r - \sigma^2/2}{\sigma}\right] - \\ - N_m\left[\frac{\ln V_f - \ln(s_2 K) + r + \sigma^2/2}{\sigma}\right]$$

Here s_2 is $> e^r$ and the sign of the expression in (18) depends on the relative size of s_2 and σ. In fact, for any $s_2 > e^r$ it is easy to see that the "riskier" the firm becomes (the larger the σ) the more (18) declines and the larger the size of the backward-looking value-maximizing capital stock relative to that of the forward-looking firm. Hence, the certainty AJ results do not necessarily hold. A similar conclusion can be derived if the value of the firm is V_b, in combination with Black-Scholes option pricing.

As a final application of the one-period valuation theory for backward-looking regulation we examine the effect of capital structure upon the value of the regulated firm. Let V_{bL} denote this value for a levered firm with an exogenously given amount of debt D. The after-tax earnings of this firm, assuming that interest and principal are tax-deductible, are equal to $\tilde{X} + t(1+r)D$. As before, V_f is the value of an unlevered firm with the same price and capital stock as the regulated firm. The previous contingent claims analysis can be applied here as well, by redefining the k_i's as $k_i \equiv \frac{s_iK - t(1+r)D - X_o}{\Delta X}$, $i = 1,2$, and noting that the value $<\tilde{X}>$ is given by the same expression as for the unlevered firm, but with the new k_i's replacing the old ones. Hence, we have

$$(19) \quad V_{bL} = tD + \frac{s_1K}{1+r} + 0(V_f, s_1K - t(1+r)D,1) - 0(V_f, s_2K - t(1+r)D,1)$$

The interesting thing in (19) is that the positive effect of leverage on the value of the regulated firm exceeds the term tD even if we neglect the effect of leverage upon the choice of optimal stock. While in forward regulation it is always true that $V_{fL}(p,K) = V_f(p,K) + tD$, by contrast $V_{bL}(p,K)$ is always $> V_b(p,K) + tD$. To see this is suffices to show that $0(S,Y_1 - \Delta,1) - 0(S,Y_2 - \Delta,1) > 0(S,Y_1,1) - 0(S,Y_2,1)$ for any pair (Y_1,Y_2) with $Y_1 < Y_2$ and any $\Delta > 0$. Since $0(S,Y,1)$ is decreasing and strictly connex in Y ([25], pp. 146) for all values of Y where $0(S,Y,1)$ is > 0, we have $-\frac{\partial 0}{\partial Y}\Big|_{Y=Y_1} > -\frac{\partial 0}{\partial Y}\Big|_{Y=Y_2} > 0$. Multiplying both sides by $d\delta$ and integrating each side from 0 to Δ we get the required inequality, Q.E.D.

V Multiperiod Extension

Any multiperiod valuation model necessitates hypothesized distributions of firm earnings beyond the first period, as well as specification of the intertemporal dependence of these distributions. Thus, in forward-looking regulation the second-period distributions of the random factors (u,W) would depend, in general on their first-period revealed values. Even if we adopt a Markovian structure of these random factors multiperiod valuation becomes quite complex analytically, given the optimal choice of output price in each period under constrained value maximization. Even under the CAPM such multiperiod valuation has, most often, adopted very simple intertemporal dependence assumptions, as in [32]. Similar difficulties appear in backward-looking regulation since the pricing rules depend on the revealed (u,W).

As in the single-period case, we assume that the initial output price p_1 is given from historical or other factors, while the capital stock K is determined optimally at the beginning of the first period. The rate of return limits s_1 and s_2 are assumed fixed, and let (u_T, W_T) denote the value of the random factors during period T. Define, for

$Q_T \equiv Q(p_T,u_T), (1-t)\ [p_T Q_T - C(Q_T,W_T,K)] \equiv X_T$

(20) $\omega_1 \equiv \{(u_T,W_T)|X_T \geq s_1 K\},\ \bar{\omega}_1 \equiv \{(u_T,W_T)|(u_T,W_T) \notin \omega_1\}$

Then the following output pricing rule under negligible regulatory lag is adopted.

(21) $p_{T+1} = p_T$ if $(u_T,W_T) \varepsilon\ \omega_1,\ p_{T+1} = \{p_{T+1}|p_{T+1}\ Q_{T+1} - C(Q_{T+1}, W_T, K) = s_1 K\}$, for $(u_T, W_T) \varepsilon\ \bar{\omega}_1$,

where in (21) $Q_{T+1} \equiv Q(p_{T+1}, u_T)$.

For multiperiod valuation purposes we shall also adopt two additional simplifying assumptions over and above those of the single period valuation. The first is that the firm operates in the region of positive marginal (with respect to price) earnings, which means that $p_{T+1} > p_T$ if $(u_T,W_T) \epsilon \bar{\omega}_1$. The second is that the earnings follow a Markovian structure given the pricing rule (21), which is equivalent to the assumption that p_{T+1} depends only on the revealed value of X_T and not on the individual values of u_T and W_T. The first assumption, in addition to its intuitive attractiveness, facilitates considerably the analysis. The second assumption is necessary for our contingent claims analysis, because it is not possible otherwise to span the space of returns of the regulated firm by primary and derivative claims on total earnings while ignoring the earnings breakdown in revenues and costs.

With these assumptions the pricing rule (21) corresponds to a Joskow-type regulation. It should be noted that (21) assumes that output price changes between periods if and only if the observed earnings X_T are below s_1K. This implies that the adjustment when X_T are above s_2K takes place by cost padding. The assumption of positive marginal earnings implies that this cost-padding is an optimal behavior when compared to output price reduction.

The valuation theory will be first developed in a two-period framework and then extended recursively to any number of periods. Let V_{bT}, T=1,2 denote the value of the backward-looking regulated firm at the beginning of period T. At the end of period 1 there are two possibilities:

either X_1 is $\geq s_1K$, in which case $p_2=p_1$, or $X_1 < s_1K$, $p_2 > p_1$, and according to the Markovian assumption $p_2 = p_2\ (X_1)$, since the solution of the equation $(1-t)[p_2Q(p_2,u_1) - C(Q(p_2,u_1), W_1,K)] = s_1K$ depends only on X_1 for all $(u_1, W_1) \in \bar{\omega}_1$. According to the notation adopted in the previous section we define $V_{fT} \equiv < X_T >^{T}_{T-1}$, T = 1,2, the value of the non-truncated earnings. However, while V_{f1} is unchanged, we have by contrast different values V_{f2} depending on whether $(u_1,W_1) \in \omega_1$ or $\bar{\omega}_1$. Let V^i_{f2}, i = 1,2 denote these two values with

(22) $V^i_{f2} \equiv [< X_2 >^2_1 | \ X_1 \geq s_1K]$, $V^2_{f2} \equiv [< X_2 >^2_1 | \ X_1 < s_1K]$. By the Markovian assumption $V^i_{f2}\ (X_1)$, i=1,2 are known functions of X_1 for given distributions of (u_2,W_2); such distributions at the beginning of period 1 are also assumed to depend only[20] on X_1.

In this framework it is easy to develop expressions for V_{b1} based on contingent claims on X_1. We need to derive the value of a contingent claim paying \$1 when $X_1 = y$, and 0 otherwise, for any $y \in [X_o, X_o + m\Delta X]$. Following [7], we do this in discrete time at first, and we start by deriving the value of a claim paying \$1 for all $X_1 \geq s_1K$, and 0 otherwise. A call option $O(V_{f1}, s_1K, 1)$ has a payoff of $X_1 - s_1K$ if $X_1 \in [X_o + k_1\Delta X, X_o + m\Delta X]$ and 0 otherwise. Similarly, $O(V_{f1}, s_1K+\Delta X, 1)$ pays $X_1 - s_1K - \Delta X$ for $X_1 \in [X_o + (k_1+1)\Delta X, X_o + m\Delta X]$ and 0 otherwise. Hence, the portfolio $C(V_{f1}, s_1K, 1) - O(V_{f1}, s_1K + \Delta X, 1)$ pays ΔX for $X_1 \in [X_o + k_1\Delta X, X_o + m\Delta X]$, and 0 otherwise. For X_1 continuously distributed, and dividing by ΔX and taking the limit we get that this contingent claims price is given by $-\frac{\partial O(S\ Y,1)}{\partial Y}$, evaluated at $S=V_{f1}$ and $Y=s_1K$. Correspondingly, since the price of a claim that pays \$1 for all X_1 is $\frac{1}{1+r}$ the value of a claim paying \$1 for $X_1 < s_1K$ and 0 otherwise is equal to $\frac{1}{1+r} + \frac{\partial O(S,Y,1)}{\partial Y}$, evaluated at the same S and Y.

By extension of this procedure it is also possible to find the value of a claim paying \$1 when $X_1 = y$ and 0 otherwise. In discrete time define $\Delta O(k) \equiv O(V_{f1}, X_o + k\Delta X, 1) - O(V_{f1}, X_o + (k+1)\Delta X, 1)$, the portfolio that pays 0 for $X_1 \leq X_o + k\Delta X$ and ΔX otherwise. Then the portfolio $\Delta O(k-1) - \Delta O(k)$ yields ΔX for $X_1 = X_o + k\Delta X$ and 0 otherwise. Dividing by ΔX twice and taking the limit as $\Delta X \to 0$ we get the pricing function $\frac{\partial^2 O(V_{f1}, y, 1)}{\partial y^2} \equiv O_{yy}$ representing the current value of a contingent claim paying \$1 for $X_1 = y$ and 0 otherwise, for all $y \in [X_1, X_o + M]$, where M is the width of the earnings range.

The value V_{b1} is now straightforward given the functions $V^i_{f2}(y)$, $i = 1,2$. Let $V^i_{b2}(y)$ denote the second-period value of the regulated firm corresponding to $V^i_{f2}(y)$, $i = 1,2$. This second-period value is given by (14), with $V^i_{f2}(y)$ replacing V_f. Hence, we have

$$(23)\quad V_{b1} = \int_{X_o}^{s_1K} O_{yy}[s_1K + V^2_{b2}(y)]\, dy + \int_{s_1K}^{s_2K} O_{yy}[y + V^1_{b2}(y)]dy + \int_{s_2K}^{X_o + M} O_{yy}[s_2K + V^1_{b2}(y)]dy$$

or, substituting for $V^i_{b2}(y)$ from (14)

$$(24)\quad V_{b1} = \int_0^{s_1K} O_{yy}[\, s_1K + \frac{s_1K}{1+r} + O(V^2_{f2}(y), s_1K, 1) - O(V^2_{f2}(y), s_2K, 1)]dy + \int_{s_1K}^{s_2K} yO_{yy}\, dy + \int_{s_2K}^{X_o + M} O_{yy}\, s_2K dy +$$

$$+ \int_{s_1K}^{X_o+M} O_{yy}\left[\frac{s_1K}{1+r} + O(V^1_{f2}(y),s_1K,1) - O(V^2_{f2}(y),s_2K,1)\right]dy =$$

$$= \frac{s_2K}{1+r} + O(V_{f1},s_1K,1) - O(V_{f1},s_2K,1) + \int_{X_o}^{s_1K} O_{yy}\, V^2_{b2}(y)dy + \int_{s_1K}^{X_o+M} O_{yy}\, V^1_{b2}(y)dy.$$

The above formulas are applicable without reformulation to any number of periods, by replacing 1 and 2 by T and T+1 respectively in every subscript of V_f and V_b.

The analytical complexity of (23) or (24) stems from the conditional structure of the second-period returns, depending on the event of the occurrence of an output price change. Conversely, if such an event is not contemplated in valuation we have a straightforward and simple extension of (15a). Assume that we have an N-period horizon, and that the random factors (u,W) are intertemporally independent. Let also $V_f \equiv V_f(p,K) \equiv \sum_{T=1}^{N} < X_T >^{T}_{T-1}$, the value of the stream of cash flows if the self-imposed upper limit did not exist. If (15a) is valid in every single-period then under backward-looking regulation each period T contributes $< X_T >^{T}_{T-1} - O(V_{fT},s_2K,T)$, where $V_{fT} \equiv < X_T >^{T}_{T-1}$ and the short option becomes a T-period European call. Hence, we have

$$(25) \quad \hat{V}_{bT} = V_f - \sum_{T=1}^{N} O(V_{fT},s_2K,T)$$

This simple generalization of (15a) may have been applicable to the "long" periods of regulatory price stability of the fifties and early sixties.

VI Discussion and Conclusions

This paper has examined the impact of Joskow-type regulatory behavior or backward-looking regulation upon the value of the firm. This impact is qualitatively different from forward-looking regulation, that has been considered almost exclusively in the financial literature until now. The difference lies in the fact that, while forward-looking regulation affects the stream of earnings of the firm only through the decision parameters, backward-looking regulation also changes the distribution of the stream of earnings for given decision parameters. This change is tied-in directly into the regulatory process itself and depends upon exogenous or self-imposed upper and lower limits.

The expressions that were developed did not depend upon any particular valuation model, since they assumed that the basic problem of single-period valuation of a random earnings flow with given decision parameters was already solved. In section III the approaches to the solution of this basic problem that have appeared in the financial literature were surveyed briefly in the context of forward-looking regulation, and their impacts upon the AJ certainty results were assessed. It was concluded that most of these results were preserved in forward-looking regulatory models.

Given now any single-period basic valuation model, the theory of arbitrage pricing through contingent claims was used to derive the single-period value of the backward-looking regulated firm. This value was then used in examining a number of well-established results for unregulated firms. It was found that backward-looking regulation changes these results in significant ways. Thus, when used in connection with value maximization, it invalidates the AJ certainty results, and it introduces a systematic bias into the MM value of the levered firm theory. Hence, it is quite possible that the large and growing empirical evidence against the AJ certainty results reflects the existence of backward-looking regulation rather then the reasons cited in these studies.[21]

In section V the value of the firm under backward-looking regulation was derived for more than one periods. Most of the resulting expressions became quite complex, and some additional assumptions were necessary for the derivation, although under the simpler backward-looking regulatory models of [33], [34] and [35] it was possible to find elegant expressions without undue restrictions. These difficulties are not peculiar to regulation, since the problem of multiperiod valuation of random income streams has been solved only under very restrictive assumptions.[22] Had similar assumptions been also adopted here it would have been possible to achieve considerable simplifications.

As a final remark it should be noted that backward-looking regulation has a number of disturbing efficiency and equity implications. Thus, the optimality of "cost-padding", already noted in [35], becomes evident. Similarly, the size of the regulator-controlled parameters s_1 and s_2 determines whether regulation confers capital gains or losses upon the firm's stockholders. Such questions were not examined in this paper, but they should form the object of further studies.

Footnotes

1) This dual role raises a well-known problem of circularity, since the the value of the rate base depends on the revenue that it generates, while these revenues, in turn, depend on the allowed rate of return and the rate base. See [18] and [39].

2) Exceptions are in [26], [3] and [22], which will be discussed in detail.

3) The earliest such model is probably the paradigm presented by Myers [31]. Systematic studies of backward-looking regulation were in [33], [34], [35] and (in a different context) [10].

4) Under the "normal" assumption that K is an essential input ($F(x_1,..,x_n,0) = 0$ for all $(x_2,..,x_n)$), $N(u,W,0) = 0$ and the equation has a solution at $K = 0$. A second solution at some $K > 0$ is guaranteed by the fact that N is concave and increasing in K.

5) Otherwise, if D is selected endogenously we reach the MM corner solution of an all-debt firm as in [30].

6) See [15] for an extended analysis of this hypothesis.

7) For instance, in [13], [14] and [19]; it is possible to show that $p_L < p$ under weaker assumptions but we shall follow the previous studies for comparison purposes.

8) See [36] and [38] for this distinction.

9) Surprisingly, although EG mention ([13], note 10) that K may be $\neq K_L$, they assume linearity of earnings or treat the two capital stocks as equal (see also [14], p. 1153). Similarly, the EG analysis is accepted uncritically in [19], p. 807.

10) Although factor-price uncertainty is not considered very often, its introduction has important mocroeconomic implications as in [37]. In financial analysis Clarke examined in [11] the results on the value of the firm of removing part of the fuel price uncertainty in electric utilities through fuel adjustment clauses. It was shown that these clauses brought in most cases a reduction in the systematic risk and, hence, a *ceteris paribus* increase in the value of the firm.

11) See [36], pp. 506-507 an extended discussion.

12) The equality of the covariances is also discussed in [19], p. 710, where it is shown to be the key to valuation under the CAPM. The separability of N(u,W,K) mentioned above is a sufficient condition for the equality of covariances.

13) See [2] for an extended discussion of these issues.

14) This is well-known in the AJ literature (see [1], [5] and [34]).

15) For an exception see [4].

16) This formulation of the spanning condition differs slightly from [12] because it follows the normalization procedure of [2] and [36].

17) See [36] for an extended discussion of the subject.

18) It should be noted, however, that in [18] it is clearly stated that their model is only an approximation to Joskow's. In [33], [34] and [35], on the other hand, the upper limit is interpreted in the spirit of the AJ certainty model. As Baron and Taggart pointed out in [3] this interpretation requires ex post lump-sum transfers of wealth from the firm to the consumers that are not observed in real life. However, these models make a lot of sense if the constraint is reinterpreted as a self-imposed upper limit in order to avoid a potentially more damaging regulatory intervention.

19) This is also true for forward-looking regulation.

20) It is possible to relax the Markovian assumption by making the distribution of (u_T, W_T) depend on the observed earnings of other firms at time T-1, but in such a case the value of the firm would depend on contingent claims of these other firms. Such an extension needs a much more complex model.

21) For a partial survey of the empirical evidence see [4], note 25. Subsequent studies that found no evidence of the AJ results were in [24] and [38].

22) For instance, in [32] under the CAPM.

FINANCING AND INVESTMENT BEHAVIOR OF THE REGULATED FIRM UNDER UNCERTAINTY

M.K. BERKOWITZ

E.G. COSGROVE

University of Toronto

Most studies of public utility regulation have concentrated on either the production decisions (e.g. Averch and Johnson (1962), Baumol and Klevorick (1970), etc.) or the financing decisions (e.g. Elton and Gruber (1971, 1977), Jaffe and Mandelker (1976), Arditti and Peles (1980), etc.). An exception to these myopic discussions is an article by Robert Meyer (1976) which attempts to integrate the "real" and financial aspects of the regulated firm. In doing so, Meyer merges the traditional weighted cost of capital concept and uncertainty with the basic Averch-Johnson model.

Unfortunately, Meyer's attempt is tainted by two fundamental misconceptions of the problem. First, it is not sufficient to simply replace the exogenously determined allowed rate of return by the weighted cost of capital in the regulatory constraint. While Meyers makes a distinction between the price of capital and the interest on funds used to finance that capital, clearly the rental price of capital includes the opportunity cost of funds so that the regulatory constraint is misspecified. Second, Meyers assumes that the correlation between the firm's returns and the returns on the market portfolio are unaffected by the firm's financing and investment decisions. It is generally accepted, however, that this systematic component of the firm's overall risk is related to the decisions undertaken by the firm. For example, Hamada (1969) and Rubinstein (1973) have demonstrated that the systematic risk is directly related to the firm's debt/equity choice. For both of these reasons, the results obtained by Meyers are subject to serious question.

In this chapter, we examine the integration of the financing and investment decisions for the regulated firm operating in an uncertain environment which includes the possibility of bankruptcy and the associated costs thereof. In doing so, our model corrects for the deficiencies in Meyer's formulation of the problem by more accurately reflecting the regulatory process and by being more consistent with generally accepted principles in finance theory. The firm must simultaneously decide on the level of capacity to employ, the method of financing, and the price to charge for its product. These decisions are compared for the regulated and unregulated firms and the impact of bankruptcy is examined within a regulatory framework in which the allowed rate of return is based upon the marginal costs of capital. These results are then compared to the decisions made under the current regulatory regime where the allowed return is based upon embedded costs. Finally, we discuss the problem of regulation within a principal-agent context in order to ascertain possible additional costs of regulation

II. Development of the Model

Consider the formation of a new firm where K in capital, which is to be determined, is needed to finance its productive activities. For simplicity, we assume no depreciation and perfect second-hand markets so that after one period, the firm receives K upon liquidation.[1] In this situation, bankruptcy does not imply the termination of the firm's activities since the firm is already assumed to be liquidated after one period. Bankruptcy occurs if, at the end of the period when the settling-up with creditors takes place, the firm lacks the necessary funds to pay its bondholders both the principal and interest owed to them.

Throughout the subsequent discussion, we use the following notation:

p : price per unit of output;

$\tilde{D}(p)$: random demand for the firm's product which is assumed to be homoskedastic and of the form, $\bar{D}(p) + \tilde{\varepsilon}$, where $\frac{\partial \bar{D}}{\partial p} < 0$;

c : constant operating cost per unit of output produced;

i : before-tax interest rate on the firm's debt;

ρ : riskless rate of interest;

B : bookvalue of firm's debt;

S : bookvalue of firm's equity;

K : capacity employed by the firm, where the purchase price per unit of capacity is normalized to equal one;

L : explicit costs associated with bankruptcy which are assumed to exceed the usual costs of liquidation at the end of the period;

δ : debt-capital ratio employed by the firm, which equals B/K ;

k^* : after-tax allowed return to equity holders of the firm;

τ : corporate tax rate;

$\tilde{\Pi}$: before tax operating profit of the firm, which is equal to $(p-c)\ \tilde{D}(p)$.

As stated above, bankruptcy occurs if the firm is unable to meet its obligations to bondholders at the end of the period, i.e. the firm is bankrupt if

$$\tilde{\Pi} + K < (1+i)B \tag{1}$$

Associated with this bankrupty condition is a probability distribution such that the cumulative probability of bankruptcy, $\text{Prob}[\tilde{\Pi}+K<(1+i)B]$, is expressed as G . The return to shareholders $(\tilde{Y}_S)$ and bondholders $(\tilde{Y}_B)$ during the period are therefore stochastic and dependent upon whether or not bankruptcy occurs.[2] Specifically,

$$\tilde{Y}_S = \begin{cases} (1-\tau)\ (\tilde{\pi}-iB) & \text{if } \tilde{\Pi}+K\geq(1+i)B \\ 0 & \text{if } \tilde{\Pi}+K<(1+i)B \end{cases}$$

$$\tag{2}$$

$$\tilde{Y}_B = \begin{cases} iB & \text{if } \tilde{\Pi}+K\geq(1+i)B \\ \tilde{\Pi}-L & \text{if } \tilde{\Pi}+K<(1+i)B \end{cases}$$

The combined returns to both shareholders and bondholders are then

$$\tilde{Y}^* = \begin{cases} \tilde{\Pi}(1-\tau) + i\tau B & \text{if } \tilde{\Pi}+K\geq(1+i)B \\ \tilde{\Pi}-L & \text{if } \tilde{\Pi}+K<(1+i)B \end{cases} \tag{3}$$

The value of the firm can be expressed under fairly general conditions by the valuation equation developed by Sharpe (1964), Lintner (1965), and Mossin (1966).[3]

$$\begin{aligned} V &= -K + \frac{E(\tilde{Y}^*) + K - \lambda\text{cov}\ (\tilde{Y}^*,\tilde{Y}_M)}{1+\rho} \\ &= \frac{E(\tilde{Y}) - \lambda\text{cov}\ (\tilde{Y},\tilde{Y}_M)}{1+\rho} \end{aligned} \tag{4}$$

where $\tilde{Y}$ is defined as the difference between Y^* and ρK, $\tilde{Y}_M$ is the return on the market portfolio, and λ is the constant price per unit of risk which is equal to $[E(Y_M) - \rho V_M]/\sigma_M^2$.[4]

While actual returns may be above or below the allowed rate of return, it is assumed that on average the firm earns the allowed return. Suppose that the competitive regulatory measured weighted cost of capital in the absence of regulation is r[5], where

$$r = \bar{k}(\frac{\bar{S}}{K}) + \bar{i}(\frac{\bar{B}}{K}) \qquad (5)$$

$\bar{k}$ and $\bar{i}$ represent the "true" opportunity costs of equity and debt funds, respectively, and $\bar{S}$ and $\bar{B}$ are the optimal competitively determined levels of equity and debt for any level of capacity (K) employed. Often in practice, $\bar{B}/K$ is referred to as the notional capital structure. In essence, r is the rental price of capital in a competitive market. In a world of certainty, the firm should be allowed to earn exactly r. However, lack of confidence in the estimates of r and the uncertainties that exist in the demand for the product suggest that a risk-averse strategy is often followed by the regulator. This results in the allowed rate being set above r in order to insure that investors earn the competitive return.

If an excess return is in fact realized, it is imputed to the shareholders so that the firm's allowed rate of return, s, can be expressed as

$$s = k^*(\frac{S}{K}) + i(\frac{B}{K}) , \qquad (6)$$

where k^* is the allowed return to shareholders. The difference between

s and r represents the upper bound of the *excess* returns per unit of capital available to the firm. It should be recognized that as the firm alters its debt-equity mix from the (regulator's) perceived optimal level of $\bar{\delta}$, the regulator may also adjust the true opportunity cost of funds in determining r so as to reflect the actual capital structure used by the firm. In the initial formulation of the problem, regulation is assumed continuous so that any change in δ that changes the opportunity cost of funds is immediately reflected in a revised allowed return on equity so as to maintain a constant (or near constant) total excess return to shareholders.

To satisfy the condition that *excess* profit does not exceed $(s-r)K$, the firm has three degrees of freedom. It chooses its product price, level of capacity, and debt-equity mix so that the following inequality is satisfied.

$$[E(\Pi) - iB]\,(1-\tau) + iB - [E(\bar{\Pi}) - \bar{i}\bar{B}]\,(1-\tau) - \bar{i}\bar{B} \leq (s-r)K, \qquad (7)$$

where $E(\bar{\Pi})$ and $\bar{B}$ represent the optimal competitive expected operating profit and debt level respectively. Substituting the values of r and s in (5) and (6), denoting the debt-capital ratio by δ , and simplifying yields

$$E(\Pi) - E(\bar{\Pi}) \leq [\alpha^*(1-\delta) - \bar{\alpha}(1-\bar{\delta}) + i\delta - \bar{i}\bar{\delta}]K \qquad (8)$$

where α^* and $\bar{\alpha}$ are $k^*/(1-\tau)$ and $\bar{k}/(1-\tau)$ respectively.[6] Implicit in the formulation of the above regulatory constraint is that any tax benefits associated with a debt-equity choice by the firm which exceeds the regulator's desired level (or notional capital structure) is regulated away through a mandated lower product price. To see this, let us look at the constraint. As long as α^* is greater than i , a marginal increase in δ above $\bar{\delta}$ results in a net reduction in the R.H.S. of the constraint equal to $(-\alpha^*+i)K$. The L.H.S. must, therefore, be similarly reduced and the firm is forced to lower the price of its product. Because, it is unclear whether or not such deviations from the perceived optimal capital structure do result in such reactive measures by the regulator, as well as the degree of the response, the regulatory constraint can more generally be expressed as

$$E(\Pi) - E(\bar{\Pi}) \leq [\alpha^*(1-\delta) - \bar{\alpha}(1-\bar{\delta}) + (i\delta - \bar{i}\bar{\delta})\phi]K \,, \qquad (9)$$

where $\phi = (1-\gamma\tau)/(1-\tau)$ and γ equals 1 when the regulator effectively regulates away all the excess tax benefits and γ equals 0 when all excess tax benefits are realized by the firm from having $\delta > \bar{\delta}$. Any value of δ between 0 and 1 represents the regulator's degree of efficiency in performing this function.

Because the firm must set its price prior to knowing actual demand, the demand for its product may exceed capacity. The costs and problems of rationing the service in the event of a shortage suggest that firms might, ex ante, choose their prices and capacity to reflect their aversion to such an occurrence. One method for dealing with this problem is the inclusion of a chance-constraint of the form.

$$\text{Prob}[\tilde{D}(p) > K] \leq \S \qquad (10)$$

The greater is the firm's aversion to unsatisfied demand, the smaller is $\S$. Such a constraint, apart from its practical relevance, has a linear equivalent form which enhances the tractability of solutions. The above inequality can be rewritten as

$$\overline{D}(p) + N\sigma_D \leq K \ , \tag{11}$$

where N is the number of standard deviations above the mean necessary to reduce the area in the upper tail of the probability distribution to § .

If the firm's objective is to maximize the market value of shareholders' wealth, it will choose p , K and δ so as to maximize the valuation expression in (4) subject to the constraints outlined in (9) and (11). Denoting the Lagragian multipliers associated with (9) and (11) by μ and v respectively, the Lagrangian expression can be formally written as:

$$\begin{aligned}\underset{p,K,\delta}{\text{Max}}\ L = \frac{1}{\rho'}\,[E(\tilde{Y}) - \lambda\text{cov}\,(\tilde{Y},\tilde{Y}_M)] + \mu[[\alpha^*(1-\delta) - \overline{\alpha}(1-\overline{\delta})\\ + (i\delta - \overline{i}\overline{\delta})\phi]K - E(\tilde{\Pi}) + E(\overline{\Pi})] + v[K - D(p) - N\sigma_D]\end{aligned} \tag{12}$$

The Kuhn-Tucker conditions for the problem outlined above are:

$$\begin{aligned}\frac{\partial L}{\partial p}:\quad E(MR) = c + \frac{1}{\psi}[v\rho' + \tau(1-G)\,\frac{\partial p}{\partial D}\,E(\varepsilon/\varepsilon \geq A) + \frac{\lambda\sigma_{DM}}{\sigma_D^2}\,\frac{\partial p}{\partial D}\,.\\ [\sigma_D^2 - \tau(1-G)E(\varepsilon^2/\varepsilon \geq A)]] + \frac{1}{\psi}\,\frac{\partial G}{\partial p}\,\frac{\partial p}{\partial D}\,X\end{aligned} \tag{13a}$$

$$\begin{aligned}\frac{\partial L}{\partial K}:\ v = \frac{\rho}{\rho'} - \frac{i\tau\delta(1-G)}{\rho'}\,[1 - \frac{\lambda\sigma_{DM}}{\sigma_D^2}\,E(\varepsilon/\varepsilon \geq A)] - \mu[\alpha^*(1-\delta)\\ - \overline{\alpha}(1-\overline{\delta}) + (i\delta - \overline{i}\overline{\delta})\phi] - \frac{\partial G}{\partial K}\,\frac{X}{\rho'}\end{aligned} \tag{13b}$$

$$\frac{\partial L}{\partial \delta} : \frac{i\tau(1-G)(1+\eta)}{\rho'} \left[1 - \frac{\lambda\sigma_{DM}}{\sigma_D^2} E(\varepsilon/\varepsilon \geq A)\right] = \mu[\alpha^* - i\phi(1+\eta)] - \frac{\partial G}{\partial \delta}\frac{X}{K} \tag{13c}$$

$$p \geq 0 , \quad K \geq 0 , \quad \delta \geq 0 , \quad \mu > 0 , \quad \text{and} \quad v \geq 0 \tag{13d}$$

where $\rho' = 1 + \rho$;

$$\eta = \frac{\delta}{i}\frac{\partial i}{\partial \delta} \geq 0 ;$$

$$\psi = 1 - \tau(1-G)\left[1 - \frac{\lambda\sigma_{DM}}{\sigma_D^2} E(\varepsilon/\varepsilon \geq A)\right] - \mu\rho' ;$$

$$X = \tau(p-c)\,[\bar{D}(p) + E(\varepsilon/\varepsilon \geq A)] - i\tau B - L + \frac{\lambda\sigma_{DM}}{\sigma_D^2}\Big[E(\varepsilon/\varepsilon \geq A)\,\Big[\Big(\frac{1-\tau G}{G}\Big)\,E(\Pi) - i\tau B - \frac{L(1-G)}{G}\Big] - (p-c)\,\Big[\frac{\sigma_D^2 - (1-\tau G)\,E(\varepsilon^2/\varepsilon \geq A)}{G}\Big]\Big] ; \text{ and}$$

$$A = \frac{(1+i)B - K - (p-c)\bar{D}}{(p-c)}$$

Equation (13a) states that the firm should set the price of its product where the expected marginal revenue equals the expected marginal cost. Notice that the expected marginal cost includes the usual operating and capacity costs as well as an explicit adjustment for risk. The greater the covariance between the demand for the firm's product and the market, the lower the price of the product.

An interesting implication of this is that a firm which sells the same product to two or more customers can "legitimately" charge different prices if the demands of the different customer classes are correlated differently with the market. For the services offered by most public utilities, this is typically the case and would suggest that differentiated pricing practices must be evaluated on <u>both</u> cost and risk differences between customer classes.

Furthermore, because X appears to be positive for reasonable values of its arguments, the third expression on the R.H.S. of (13a) depends upon the change in the cumulative probability of bankruptcy as the price increases. If we assume that $\tilde{\varepsilon}$ is normally distributed, this change in the cumulative probability becomes:

$$\frac{\partial G}{\partial p} = -g(A) \left[\frac{(p-c)^2 \frac{\partial \bar{D}}{\partial p} + (1+i)B-K}{(p-c)^2}\right] , \tag{14}$$

where g(A) is the marginal probability of A . It should be recognized, however, that the theory suggests that the probability distribution is not continuous over the entire range, but instead, the probability of bankruptcy is discontinuous at some critical δ , say δ^* , at which point the probability of bankruptcy increases substantially for values of δ above δ^* . The above expression can therefore be thought of as an approximation of the actual distribution in order to explicitly show the variables affecting $\frac{\partial G}{\partial p}$, For values of δ below δ^* , it should be realized, however, that $\frac{\partial G}{\partial p}$ is negligible while for values of δ greater than or equal to δ^* , $\frac{\partial G}{\partial p}$ is indeed significant.

Looking to (14), we see that the sign of $\frac{\partial G}{\partial p}$ is related to the demand elasticity, interest and debt principal, and total capacity expenditure. The sign of these relevant variables suggests that $\frac{\partial G}{\partial p}$ is positive. Therefore, it follows that the third expression in (13a) is positive though small for $\delta<\delta^*$. This means that a higher price, which increases the risk of bankruptcy, reduces prices yet further than they would have been in the absence of the threat of bankruptcy.

Condition (13b) describes the firm's capacity decision. Capacity is added to the point where the expected marginal contribution of the last unit employed exactly equal its expected marginal cost. It is interesting that when looking at (13b), we see that the marginal capacity cost is the sum of the discounted riskless rate and an adjustment for risk which is related to the correlation of the demand with the market (as expressed in the second part of the second term). The effect of regulation, which is assumed binding $(\mu>0)$, is the not surprising result that more capacity is employed relative to an unregulated, but otherwise identical firm. Moreover, the fourth expression on the R.H.S. of (13b) is positive since it can be easily demonstrated that $\frac{\partial G}{\partial K}$ is negative,[7] though again quite small for $\delta<\delta^*$. As K increases and the probability of bankruptcy subsequently decreases, there appears to be a negative consequence of such behavior that can be seen by examining the components of X. While the firm loses the tax benefits from its debt and must incur the explicit costs of bankruptcy in the event of such a mishap, it also will _not_ pay taxes on its _ex ante_ expected income as it had planned to do when it made its capacity decision at the beginning of the period. Although this is unquestionably a small effect when δ is less than the critical value; this opportunity benefit is an inducement to a somewhat lower level of capacity, below that already distorted level induced by regulation.

What we have referred to above as an opportunity benefit of bankruptcy has typically been overlooked when the costs of bankruptcy have been examined in the literature. For example, Kim (1978) states that bankruptcy costs can be thought of as being comprised of three major components. First, there is the "short-fall" arising from liquidation or the "indirect" cost of reorganization, both of which are absent in a single-period model such as the one being analyzed in this paper. Second, various administrative expenses must be paid to third parties in the course of the bankruptcy proceedings, represented by L in our model. Third, firms lose tax credits which they would have received had they not gone bankrupt, as expressed by $i\tau B$. In addition, however, there is the tax that was expected to be paid, but in the event of bankruptcy, will not be paid, and is represented as $\tau(p-c)[\overline{D}(p) + E(\varepsilon/\varepsilon \geq A)]$. This latter expression can be though of as a negative cost (or benefit) of bankruptcy.

The final condition (13c), aside from the non-negativity conditions summarized in (13d), describes the firm's debt-equity choice. This equation suggests that the optimal debt-equity mix is the one in which the discounted expected marginal benefits from an extra dollar of debt exactly equals the expected marginal cost associated with that dollar of debt. The expected marginal cost of debt, as expressed by the R.H.S. of (13c), includes two terms. The first denotes the opportunity cost of debt since the cost of equity exceeds the debt cost. Therefore, as debt is substituted for equity the allowed return is reduced. Notice that as the elasticity of the interest rate increases, the opportunity cost of having additional debt is reduced as long as the excess return to shareholders remains constant (i.e. regulation is continuous). The greater is the elasticity in this situation, the smaller is the financing inefficiency.

The degree of reduction in the opportunity cost depends as well on the extent to which the tax benefits from debt are regulated away through lower product prices. In the extreme case where γ is equal to one, implying all the tax benefits from excess debt are regulated away, ϕ equals one. At the other extreme, when γ equal zero, implying all the tax benefits from excess debt are realized by the firm, ϕ equals $(1/1-\tau)$. Because $(1/1-\tau)$ exceeds 1, the marginal cost of debt is lower when all tax benefits from debt are allowed to be realized, ceteris parabis, and the firm will choose a higher debt-equity mix than otherwise.

The second term on the R.H.S. of (13c) is negative since $\frac{\partial G}{\partial \delta}$ is positive.[8] That is, because an increase in the debt-equity ratio increases the probability of bankruptcy, there is a higher probability that the <u>ex ante</u> expected tax payment will not have to be made. Thus, the increased risk of bankruptcy has a small (for $\delta<\delta^*$), but positive effect on the firm's decision to employ more debt.

Before, leaving our discussion of the first-order conditions, we should mention that these results are quite general as well as robust. For example, suppose we assume that the regulatory constraint is not binding (i.e. $\mu=0$) and bankruptcy cannot occur. Then condition (13c) implies that in equilibrium the marginal benefit from an extra dollar of debt is zero which suggests 100 percent debt financing —

the well known result arrived at by Modigliani and Miller (1963). Once the threat of bankruptcy is introduced, again assuming $\mu=0$, and the bankruptcy costs (L) are large enough to cause X to be negative, a finite level of debt is dictated by (13c) even in the presence of corporate taxes. That is, if the bankruptcy costs are large and the probability of bankruptcy takes a discontinuous jump at δ^*, the optimal mix will be marginally below δ^*. In a somewhat different manner, Stiglitz (1969) arrived at the same result in the presence of bankruptcy and taxes.

Finally, let us examine the debt-equity choice for the regulated firm relative to an otherwise identical but unregulated firm. If we assume for the moment that the second term on the R.H.S. of (13c) is constant, irrespective of the firm being regulated or not, then it appears that the unregulated firm ($\mu=0$) will employ more debt than its regulated ($\mu>0$) counterpart. It is likely, however, that the second term is not constant. If we interpret the cost of bankruptcy, L, as including lost earnings in the event of bankruptcy,[9] clearly the cost of bankruptcy for the unregulated firm exceeds that for the regulated firm. Therefore, the smaller the excess return allowed the regulated firm, the greater will be the relative loss in earnings for the unregulated firm. It is possible, moreover, that L is so large for the unregulated firm, as compared to the regulated firm, that X will reverse in sign to become negative so that the R.H.S. of (13c) for the unregulated firm exceeds the R.H.S. for the regulated firm. If this occurred, the impact of regulation would be opposite to the earlier result, i.e. the unregulated firm would have a lower optimal debt-equity mix than it would if it were regulated. This conclusion was reached by Arditti and Peles (1980) who argued that the firm has less to lose when it is regulated and will therefore issue a greater amount of debt than the unregulated firm which has more to lose if bankruptcy occurs. While one is unable to unequivocably state the direction of the regulated and unregulated debt-equity levels in this model, we are able to discern the relevant factors and appreciate the complexity of this problem, something avoided in the simplified world envisioned by Arditti and Peles.

III. Regulation Based Upon Embedded Costs

When the regulatory constraint is formulated on an embedded cost basis, as is typically the case, the component costs of debt and equity are weighted averages of the outstanding and proposed issues. In this situation the firm enters the period with an accumulated stock of capital, $\hat{K}$, which has been financed by debt and equity of $\hat{B}$ and $\hat{S}$, respectively. Because we continue to maintain that excess returns accrue only to shareholders, and the excess return is constant, it is sufficient to compare the debt costs under the two regulatory regimes in order to evaluate the effect of marginal cost-based regulation.

The embedded cost of debt, i', can be represented as

$$i' = \frac{\hat{i}\,\hat{B}}{\hat{B} + B} + \frac{i\,B}{\hat{B} + B} \qquad (15)$$

The greater the increase in interest rates during the past years and the greater the proportion of the firm's outstanding debt that has been financed at the previous lower rates, the smaller that i' is relative to i . Once this embedded cost of debt is substituted into the regulatory constraint, the revised constraint becomes:

$$E(\Pi) - E(\overline{\Pi}) \leq [\alpha^*(1-\delta) - \overline{\alpha}(1-\overline{\delta}) + (i\delta - \overline{i}\overline{\delta})\phi]\,(\hat{K}+K) \qquad (16)$$

Without formally presenting the first-order conditions associated with the amended problem, it is fairly straightforward to compare the results to those in (13a)-(13c). Because the embedded cost of debt is below the marginal cost, equation (13c) would dictate a greater substitution to equity from debt in order that the allowed return might be increased. The tradeoff between a higher regulated return and increased tax benefits is even more one-sided in favor of the former incentive than in the earlier problem. Furthermore, it is quite easy to show that upon solving (13c) for μ , the increase in allowed return, due to a lower interest rate on debt in the regulatory constraint, reduces the benefit, μ , from relaxing the constraint. In turn the lower μ and δ have the effect in (13b) of increasing the expected marginal capacity cost so that less capital is employed relative to marginal cost-based regulation. In essence, what embedded cost regulation has done compared to marginal cost regulation is to induce the firm to substitute greater financial inefficiency for less production inefficiency. Unfortunately the effect of this substitution upon the product price is directly related to the relative sizes of the inefficiencies and their respective effects can only be determined by assuming specific functional forms.

IV. Agency Costs of Regulation

Throughout our discussion we have consistently referred to the firm's decisions where it has been implicitly assumed that the managers of the firm always act in the best interests of the shareholder-owners, or to be even more narrowly specified, the managers are the owners of the firm. The literature on regulation has followed a similar path. Though the divergence of interests between the managers and owners has been recognized, the resulting consumption of perquisites by managers has been overlooked for one reason or another. This apparent oversight has, however, been redressed in the general economics literature where the incentive problem has received a great deal of attention.[10] The agent-principal problem has also been discussed with respect to its impact on the firm's financing and investment decisions - most notably by Jensen and Meckling (1976).

In the Jensen-Meckling model, an agency problem arises from the fact that with a fixed money wage, a manager who owns less than 100 percent of the firm's stock, say κ , imputes to himself only the fraction κ of the lower value of the stock when he consumes more on the job. As a consequence, the manager consumes more shirking and perquisites the smaller is his fraction of ownership of the firm's common stock. It should be recognized that perquisites may take the form of (suboptimal) decisions which are not consistent with the interests of shareholders. Moreover, in the J-M model, monitoring costs are assumed to vary inversely with the fraction of the firm owned by the manager(s). The manager has, therefore, an even stronger incentive to shirk and consume perquisites the smaller is his ownership share because his consumption is more costly to monitor and control. As a result, the optimality of the investment package and method of financing the investment varies inversely with κ .

While the costs of inefficiency due to regulation on both the production and financing sides have been examined, the effect of regulation on the size of the agency costs within the firm has thus far not been discussed in the literature. To appreciate the implications of regulation in this light, it is first necessary to examine the role of regulation in the principal-agent relationship.

In the absence of regulation, the firm can be described by the usual principal-agent relationship. The manager chooses a set of actions and then shares the consequences of the actions with the principal. In performing his managerial functions, the interests of the manager are not always consistent with those of the owners so that costs are incurred which are directly attributable to this relationship. To reduce these costs, the owners can institute monitoring at some cost. The greater the expenditure on monitoring, presumably the lower are the agency costs, i.e. consumption of perquisites and shirking.

Because of the natural monopoly characteristics of these firms, regulation has most often been suggested as the remedy that allows the firm to operate at the scale of a monopolist yet not charge prices which reflect that degree of monopoly power. In this regard, the regulator is the watchdog of the public interest - i.e. consumers of the product and owners of the firm alike.

With the addition of regulation, the principal-agent relationship becomes more complex. The managers of the firm are no longer simply responsible to the owners, but must now satisfy the interest of consumers as well. The function of regulation in this respect is two-fold. First, it monitors the decisions of the manager so as to insure that the actions taken are in the best interests of the co-principals (consumers and owners) jointly. Because the agent now has a responsibility to two principals with separate interests, the actions taken by him are suboptimal from the standpoint of each principal. This has been witnessed in our earlier discussion of the induced inefficiencies ascribed to the regulatory process. The second function of the regulator is to provide an equitable distribution between the principals of the outcome of the manager's actions. This is the dynamic aspect of regulation. If the manager in the present period accepts a project and next period earns excess rents on that project, the regulator must decide the portion of those rents which should be imputed to consumers by way of lower prices and what portion should be imputed to shareholders.

Our concern is whether the managers of a regulated firm pursue actions so as to incur agency costs which are directly attributable to the process of regulation. While it is recognized that regulation itself may have its costs, both administrative and inefficiency, these are not agency costs since the manager was always implicitly assumed to operate in the owner's interest. Given the regulatory environment in which the manager operates, overcapitalization, for example, would be in the interest of shareholders.

While it is not the purpose of this discussion to identify all the possible costs of the agency relationship which are attributable to regulation, one example demonstrates that an evaluation of current regulatory practices should consider these additional costs. For our particular example, it is useful to examine the proceedings of a regulatory review. Typically, the managers, purportedly acting in the interest of the owners, argue that the allowed return be increased for one reason or another. Consumer-intervenors, on the other hand, vehemently argue that the allowed return should be lower than the rate suggested by the managers. Clearly, the managers are in a dubious position. They realize that part of their performance as managers is to achieve as high an allowed return as possible for the owner-shareholders. Yet, having been successful and being allowed a higher return, they must adopt investments of higher risk in order to realize the high return. Because managers are generally perceived as being risk averse and recognize that such an investment policy could jeopardize their position in the firm, they have a choice of actions. On the one hand, their arguments for an increased allowed return can be presented in a less than optimal manner so that only a marginal increase is allowed by the regulator. In this situation, they are able to make relatively less risky investments and achieve the allowed rate with little additional risk to themselves. On the other hand, they may argue and achieve a higher allowed return after which they adopt investments of a risk which assures that their realized return is below the allowed level. In both of these situations there are agency costs. The action taken by the manager is the one which is more difficult (or expensive) to monitor. It appears in this case that the manager will choose to act suboptimally during the regulatory review since the alternative action is quite easy to monitor.

As we stated above, we are not attempting here the task of identifying all the possible areas in which regulation might affect the costs of the agency relationship. Yet we should point out in closing that an avenue worthy of future research is the effect of regulation on the risk structure of the firm and the consequential actions by risk-averse managers in response to this change in risk due to regulation.

V. Summary and Conclusions

In this chapter we have examined the integration of the financing and investment decisions for the regulated and unregulated firm operating in an uncertain environment which includes the prospect of bankruptcy. Contrary to current regulatory practice which bases the allowed return on embedded costs, our model postulated a regulatory framework in which the allowed return was based on marginal costs and was therefore consistent with the investment decision. The model developed was shown to be both general and robust in that it is capable of demonstrating the early theories of capital structure much discussed in the finance literature. Furthermore, it was possible using this model to show that regulation induces both a production and financing inefficiency. Despite popular belief, it is likely that the regulated firm operates with too little debt. Moreover, when we compared the financing decisions for the regulated and unregulated firms, we were able to identify the relevant parameters and their respective magnitudes necessary for an unambiguous comparison.

The optimal decisions under marginal cost based regulation were then compared to those under embedded cost regulation. In doing so, it was shown that these different forms of regulation lead to different tradeoffs between the inefficiencies in investment and the method of financing the investment. The magnitude of the total cost of inefficiency associated with each method of regulation must be compared in order to identify the least inefficient practice.

Finally, the problem of regulation was characterized within the principal-agent relationship. Within this framework, we discussed the possibility of additional agency costs which can be directly attributable to the regulatory process. While regulation has consistently throughout the years been attacked for the inefficiencies which it causes, it may very well be that the additional costs of the agency relationship, which are attributable to regulation, are of sufficient order to further rebuke any benefits ascribed to the regulatory process. At the least, it is a subject worthy of further pursuit.

Footnotes

1. If depreciation (obsolesence) occurred at some rate d throughout the period, the firm would realize $(1-d)K$ at the end of the period. The per unit cost of capacity in (4) would then be $(\rho+d)$ instead of ρ. This additional consideration would effect the bankruptcy condition in (1) and our results would change accordingly. Moreover, if the value of K was uncertain at the end of the period, the certainty equivalent end-of-period value of the capital would appear in (4) and the results again would be reinterpreted.

2. We have simplified the model by assuming that in the event of bankruptcy the firm (bondholders) does not have to pay taxes. In a multi-period setting, the firm could suffer several period of losses or low profitability without being forced into bankruptcy as long as interest obligations (and any other bond covenants) were fulfilled. Since these losses are quite common in years preceeding the actual bankruptcy, the firm would probably be able to carry forward any previous losses and eliminate any tax liability that might arise should the firm show a "taxable profit" in its final period.

3. Although the Capital Asset Pricing Model (CAPM) is necessary to provide a simple, but practical method for examing the problem of optimal capital structure, one does not need to assume CAPM. Instead, a more general theoretical model such as the state preference approach could be adopted, in which case it is likely that the additional complexity of the model will greatly inhibit implementation.

See, for example, Kraus and Litzenberger (1973) who suggest within a state preference framework that a stochastic dynamic programming approach should be used to search for an optimal capital structure.

4. After simplifying the expression,

$$E(\tilde{Y}) = (p-c)D[1-\tau(1-G)] - \rho K + i\tau B(1-G) - LG - \tau(p-c)(1-G)E(\varepsilon/\varepsilon\geq A)$$

and

$$\mathrm{Cov}(\tilde{Y},\tilde{Y}_M) = \frac{\sigma_{DM}}{\sigma_D^2}[[(1-G)E(\varepsilon/\varepsilon\leq A)(1-\tau) + GE(\varepsilon/\varepsilon\geq A)]E(\tilde{\pi}) + (1-G)E(\varepsilon/\varepsilon\geq A)i\tau B - GLE(\varepsilon/\varepsilon\leq A) + (p-c)\cdot [(1-\tau)(1-G)E(\varepsilon^2/\varepsilon\geq A) + GE(\varepsilon^2/\varepsilon\leq A)]] ,$$

where $A = \frac{(1+i)B-K-(p-c)\bar{D}}{p-c}$

The term $E(\varepsilon/\varepsilon\geq A)$ represents the expected value of ε conditional upon bankruptcy not occurring, i.e. $\varepsilon\geq A$. Similarly, $E(\varepsilon^2/\varepsilon\geq A)$ is the variance of ε given bankruptcy does not occur.

5. The practice in regulatory proceedings is to measure the overall return, and hence the cost of funds in the absence of transaction costs, by weighting the sum of the after-tax return on equity and before-tax return on debt. We refer to this sum as the regulatory measured weighted cost of capital, as distinguished from the generally accepted cost of capital which is a weighted sum of the after-tax costs of both equity and debt.

6. Though equity costs are measured after-tax and debt costs before-tax in order to reflect actual regulatory behavior, the inequality in (8) is fully consistent since both sides are presented on a before-tax basis.

7. Again, assuming $\tilde{\varepsilon}$ is normally distributed,

$$\frac{\partial G}{\partial K} = g(A)\left[\frac{(1+i)\delta-1}{(p-c)}\right]$$

which is greater than, equal to, or less than zero as $(1+i)\delta \gtrless 1$. For reasonable values of δ , it follows that $\frac{\partial G}{\partial K} < 0$.

8. It follows from our assumption of $\tilde{\varepsilon}$ being normally distributed that

$$\frac{\partial G}{\partial \delta} = g(A)\left[\frac{(1+i)K}{(p-c)}\right] > 0 .$$

9. It is possible, furthermore, to assume that L is stochastic within the model, where $\tilde{L} = 0$ if bankruptcy does not occur and $\tilde{L} = L(\tilde{\pi})$ if bankruptcy occurs. The specification was adopted by Kim (1978).

10. One of the earliest reasons given for the incentive problem that exists between the principal and the agent was a difference in risk attitudes held by the two parties. Within this context, Arrow (1971) and Wilson (1968) examined the optimal sharing of purely exogenous risk. Later, Wilson (1969) and Ross (1973) considered situations in which risk could be affected by the actions of the agent. In contrast, Spence and Zeckhauser (1971) analyzed the problem of divergency in incentives as a result of differential information, which was subsequently extended by Harris and Raviv (1979). Following this, Shavell (1979) and Holmstrom (1979) have examined the problems associated with imperfect monitoring.

TAXES, FINANCING AND INVESTMENT FOR A REGULATED FIRM

JEFFREY I. BERNSTEIN

McGill University

1. Introduction

This paper develops a model of integrated investment and financing decisions for a regulated firm. The first objective is to characterize the determinants of corporate financial policies and impacts which emanate from regulation. Second, we desire to show, given the intertemporal nature of the problem, how corporate policies change over time both within and across regulatory regimes.

The theory of investment has been well established and clarified since the early 1970's by Lucas [11] and Treadway [15], among others. However, the analysis of investment and regulation has been given quite limited attention. Two notable exceptions are the papers by Appelbaum and Harris [1] and Klevorick [10], which like the traditional static model (as described in Baumol and Klevorick [3]) focus on rate of return regulation.

The theory of financial decisions for rate of return regulated firms as exemplified by Elton and Gruber [5], [6] and Jaffe and Mandelker [7], although setting the stage for a more complete analysis, has neglected the intricacies of corporate investment decisions, and an analysis of the alternative intertemporal financing patterns.

The problem of investment, financing and taxation for unregulated firms was set out by Stiglitz [14] and then later expanded by King [9]. The focus of this literature largely concentrated on the manner in which personal and corporate taxes affect the user cost of capital. Stiglitz showed how the tax system was neutral with respect to the user cost. However, King by introducing the institutional constraint that dividends cannot be paid from new bond and share issues, pointed out the distortion to the corporate equilibrium.

The dynamic model constructed within this study is of a firm which is able to finance investment through retentions, bonds, which can have various terms to maturity, and shares. Accompanying investment, adjustment costs are incurred by the firm, which are external to the production process and based on gross physical investment. Regulation is not of the rate of return variety, but appears as a limit on corporate earnings. In particular, there is an upper bound on the addition to accumulated earnings that the firm can earn in any period, while it is being subject to continuous regulatory review.

Limiting the rate of return has been criticized (for example, by Joskow [8] and Panzar and Willig [13]) as not being reflective of the constraining nature of regulation. This criticism is relevant if we select the Canadian Radio-television and Telecommunications Commission (CRTC) as our representative regulatory agency. A casual observation of the manner in which the CRTC regulates the telecommunications carriers under its jurisdiction, demonstrates the weakness of the rate of return paradigm. The CRTC establishes the operating expenses, the investment programme (through the construction programme review committee), the rate of return, the quantity and quality of service (for example, the non-urban service improvement package), and analyses the determinants of demand, in order to come to a conclusion concerning revenue requirements and price structure. In addition the commission approves new bond and share issues.

Analytically within a static context rate of return regulation may not be a bad approximation (perhaps even for a dynamic model which ignores the financial questions). But when one becomes interested in studying the problem of investment and financing, rate of return regulation is quite limited. First, the regulatory constraint does not contain any corporate decision (or control) variables. This implies that the short run equilibrium is not affected by changes in the allowed rate of return (i.e. in the decisions of the regulator). Second if the allowed rate of return is defined net of adjustment costs, then the firm does not have any choice in its investment programme. Given the initial capital stock (and the associated financing pattern) and the allowed rate of return, investment is completely determined. This is true irrespective of the corporate objective. However, we do know that regulated firms are affected in the short run by the decisions of the regulator, while at the same time the regulator, itself, does not set the investment path of the firm.

Finally, rate of return regulation does not permit us to explain the salient differences in financing patterns between regulated and unregulated firms. Regulated companies exhibit more stable dividend policies and issue shares and bonds more frequently. It is impossible, of course, to capture all of the complexities of the structure and behaviour of participants in the regulatory arena in a single model. Nor for that matter is it necessary, if we are dealing with the specifics of investment and financing. In a general sense the regulator is concerned with the level of economic profits which, in an environment where financing is endogenous, translates into a constraint on the level of retained earnings.

Corporate financial policy is determined by a comparison between the post tax marginal cost of corporate debt and the post tax marginal cost of personal debt for the shareholders, after taking into consideration the tax savings from receiving income in the form of capital gains rather than as interest. The unregulated firm finances investment through internally generated funds, with any requirements in excess of retentions financed through new share issues if the post tax marginal cost of personal debt is less than for corporate debt,and new bonds for the converse. A regulated firm because of the limitation on retentions, will divert funds to dividends in order to satisfy the constraint and thereby to a greater extent turn to the financial capital markets. We expect, if debt is the financial instrument, that investment and physical capital are larger relative to the unregulated firm. However, when new shares are offered (as opposed to bonds) the converse occurs. The reason stems from the interest deductibility provision. Although this provision does not play a role in setting the financial policy, once debt is found to be the cheaper instrument, interest deductibility permits a larger capital stock. The all equity case forces the firm to retire its debt, and given the regulated limit on retentions, funds are diverted from the expansion of its capital stock through investment.

2. The Model

Consider a corporation which operates in n non-capital input and output markets and a single physical capital market. The non-capital factor and physical capital markets are competitive and at least one of the product markets is monopolistic. The technology is represented by

$$\mathcal{P}(y(t),\ K(t),\ t) = 0$$

where $y(t)$ is the n dimensional vector of outputs and non-capital inputs, $K(t)$ is the physical capital and t (time) represents technological change. $\mathcal{P}$ is twice continuously differentiable in $y(t)$, $K(t)$, non-decreasing and strictly quasi-concave in $y(t)$.

Let the first m elements in the vector $y(t)$ be the products which are monopolistically produced. Defining $p(t)$ as the price vector for the products and non-capital inputs, then we assume that $p_i(t) = \mathcal{D}_i(y_i(t))$ $i=1,\ldots,m$ with $\mathcal{D}_i$ the twice continuously differentiable inverse demand function and $\mathcal{D}_i'<0$, while $p_j(t)$ $j=m+1,\ldots,n$ are exogenous to the firm.

At any time period the firm, given the physical capital, selects $y(t)$ by maximizing $\sum_{i=1}^{m} \mathcal{D}_i(y_i(t))y_i(t) + \sum_{j=m+1}^{n} p_j(t)y_j(t)$ subject to $\mathcal{P}(y(t),K(t),t)=0$. With $\mathcal{D}_i(y_i(t))y_i(t)$ strictly concave in $y_i(t)$ $i=1,\ldots,m$, the solution to this problem can be denoted as $y(t) = g(K(t),t)$ (with p_j $j=m+1,\ldots,n$ suppressed). Thus the maximized value of variable profits or the indirect variable profit function is

$$\mathcal{R}(K,t) = \sum_{i=1}^{m} \mathcal{D}_i(g_i(K(t),t))g_i(K(t),t) + \sum_{j=m+1}^{n} p_j g_j\ (K(t),t),\text{ where}$$

$\mathcal{R}$ is twice continuously differentiable in $K(t)$, $\mathcal{R}_k>0$, $\mathcal{R}_{kk}<0$.

The indirect variable profit function represents the flow of funds as revenues to the firm and from the firm for payment to the non-capital factors of production. The remaining flow of funds pertains to the transactions involving physical and financial capital. Equation (1) characterizes the sources and uses of corporate funds;[1]

$$(1)\qquad 0 = \mathcal{R}(K_t,t) - A(I_t) - r_{bt}B_t - T_t + b_t - \psi B_t + s_t - p_{It}I_t - D_t$$

where I_t is physical investment, A is the twice continuously differentiable adjustment cost function (see Lucas [11] and Treadway [15]) $A' \gtreqless 0$ as $I \gtreqless 0$, $A''>0$; r_{bt} is the interest rate on corporate debt, B_t is corporate debt, T_t are corporate taxes, b_t is the nominal value of new debt, $0<\psi\leq 1$ is the constant proportion of debt retired in any period, s_t is the nominal value of new shares, p_{It} is the price of physical investment, and D_t are dividends.

In this model we do not assume that all debt has a term to maturity of a single period. Corporate debt can have different maturity dates. However, to avoid the complications of a term structure problem, we assume that all debt issued at a particular date has the same term to maturity. In period t (for example) the new debt has a maturity date of t + n periods later, while the debt issued in $\mathcal{t}$ has a maturity date of $\mathcal{t}$ + m. This means that r_{bt} (since it is variable over time) is a weighted average of interest rates.

The weights equal the proportion of debt maturing in any specific period relative to the total outstanding debt.[2] With the issuance and retirement of debt given in equation (1), the net debt accumulation is

$$(2) \qquad \dot{B}_t = b_t - \psi B_t \quad , \quad B_0 > 0 \ .$$

The value of net debt accumulation depends on whether new debt issues exceed, equal or fall short of retirements, in any time period.[3] Moreover, because we are interested in the financial and physical decisions of a non-financial corporation, we assume that $b_t \geqq 0$. Thus although net debt changes can be negative, the firm itself does not demand corporate debt.

Corporate taxes are

$$(3) \qquad T_t = u_{ct}[R(K_t,t) - A(I_t) - r_{bt}B_t - \delta p_{It}K_t]$$

where $0<u_{ct}<1$ is the corporate income tax rate, $0<\delta\leqq 1$ is the fixed rate of depreciation.

Substituting Equation (3) into (1) and solving for dividends yields,

$$(4) \qquad D_t = (1-u_{ct})[R(K_t,t) - A(I_t) - r_{bt}B_t] + u_{ct}\delta p_{It}K_t + b_t - \psi B_t + s_t - P_{It}I_t \ .$$

By their very nature dividends cannot be negative, so $D_t \geqq 0$. In addition, corporations because of legislative and securities regulations must meet certain requirements in order to pay out dividends to their shareholders. In our model dividends cannot exceed the net flow of funds excluding those associated with new capital. In other words

$$(5) \qquad D_t \leq (1-u_{ct})[R(K_t,t) - A(I_t) - r_{bt}B_t] + u_{ct}\delta p_{It}I_t - \psi B_t \quad .$$

The previous inequality implies, from Equation (4), that $p_{It}I_t - b_t - s_t \geq 0$ or that the value of new debt and share issues must go towards the financing of physical investment and not for the payment of dividends. Thus we can define $E_t = p_{It}I_t - b_t - s_t$ as retentions, which by (5) must be nonnegative, and $D_t + E_t = F_t$ where F_t is the right side of (5).

In many economies there are restrictions on the ability of firms to repurchase their shares. Part of the reason is the asymmetrical treatment of dividend income and capital gains by the tax authorities. In a repurchase situation the proceeds received by the shareholders would be taxed at the capital gains tax rate which is less than the rate for dividend income. However, regular repurchase of shares would be construed as equivalent to dividends for tax purposes. Intermittent repurchases would presumably avoid this, but would subject directors and officers to the risk of shareholder suits based on the grounds that they benefitted from insider information in deciding when the firm should repurchase and whether they should sell their shares at that time. Thus in the present model we assume that the corporation cannot repurchase its shares, $s_t \geq 0$. Notice that with $E_t = p_{It}I_t - b_t - s_t \geq 0$, $b_t \geq 0$ and $s_t \geq 0$, then $I_t \geq 0$. Physical investment is irreversible because retentions, new debt and shares issues are nonnegative.

Regulation is imposed as an upper limit on the post tax earnings. In particular, we assume that the regulatory authorities restrict retentions. One way to formalize this constraint is in terms of a retention to asset ratio,

(6) $E_t \leqq i p_{It} K_t$.

where i is the allowed retentions to asset ratio. Since $E = p_{It} I_t - b_t - s_t$, then from (6) there is a limit on the ability of the firm to finance investment from internally generated funds. Moreover, by combining (5) and (6) we find that $0 \leqq E_t \leqq i p_{It} K_t$. Clearly with a positive capital stock, either the constraint denoted by (5) is effective or the regulatory constraint is binding.

In the present context i is independent of time. With no difficulty at all, we could assume that the allowed ratio varied with the time period. However, given that a regulatory authority is altering i, it would appear to be important to specify the relationship determining the movement in the allowed rate.[4] In the present context we assume away the problem of regulatory lag and treat i as time independent.

The firm is owned by a single class of shareholder, whose non-capital gains income is taxed at the rate $0 \leqq u_{pt} < 1$, and capital gains are taxed at the rate $0 \leqq u_{gt} < 1$.[5] The present institutional setting is governed by $u_{gt} < u_{pt}$. Shareholders will purchase shares to the point where the marginal after tax return from a dollar is equalized across all investments. Letting r_t be the interest rate on the alternative investment to the corporate issues, the equilibrium condition is

$$(7) \qquad (1 - u_{pt})\, r_t p_{st} S_t = (1 - u_{pt}) D_t + (1 - u_{gt})\, \dot{p}_{st} S_t \,.$$

The left side of (7) is the post tax return on investing in the alternative instrument, the equivalent dollar value that is invested in the corporation's shares. The right side is the post tax return from purchasing $p_{st}S_t$ value of shares. The shareholders receive $(1 - u_{pt})$ on every dollar paid out as dividends per share and receive $(1 - u_{gt})$ on every dollar change in the price of the share. Manifestly, the shareholders operate in a certain environment or are risk neutral.

By defining the market value of shares as $V_t = p_{st}S_t$, with $\dot{V}_t = \dot{p}_{st}S_t + p_{st}\dot{S}_t$ where $s_t = p_{st}\dot{S}_t$, then we can rewrite (7) as

$$(8) \qquad \dot{V}_t - [1 - u_{pt})/(1 - u_{gt})] r_t V_t = - [(1 - u_{pt})/(1 - u_{gt})]$$
$$[D_t - s_t(1 - u_{gt})/(1 - u_{pt})] \,.$$

Equation (8) can be solved for the initial market value of shares,

$$(9) \qquad V_0 = \int_0^\infty e^{-\int_0^t \rho_z dz} (D_t - s_t/a_t)dt$$

without loss of generality $a_0 = 1$, and we define $a_t = (1 - u_{pt})/(1 - u_{gt})$ and $\rho_t = a_t r_t - \dot{a}_t/a_t$.[6] The initial market value, given by Equation (9), depends on the dividends derived from ownership minus the pre tax dilution from the issuance of new shares. The value of this difference is discounted by the post tax return on a dollar invested in the alternative instrument. By post tax, we mean that the individuals have been compensated for the impact of capital gains which accrue to shareholders and not bondholders. The fact that both personal income and capital gains tax rates change through time, implies that present corporate decisions are affected by shareholder expectations concerning the future values for those rates.

Besides the irreversibility of investment, we assume that capital accumulates in the usual fashion,

$$(10) \qquad \dot{K}_t = I_t - \delta K_t,\ K_0 > 0$$

with $0 < \delta \leq 1$ as the fixed rate of depreciation.

The firm selects physical investment, new debt and share issues which maximize the initial market value of the shares subject to the constraints given as $s \geq 0$, $D \geq 0$, $b \geq 0$, and $0 \leq E \leq i p_I K$. The solution to this optimal control problem is characterized by the following equations, with $\phi = 1 - 1/a$, q_1, q_2, are the costate variables associated with K, and B; λ's are the Lagrangean multipliers, with λ_1 to λ_5 associated with $s \geq 0$, $E \leq p_I K$, $D \geq 0$, $b \geq 0$ and $E \geq 0$.[7]

$$(11.1) \qquad -(1+\lambda_3)(1-u_c)A' - p_I(1+\lambda_2+\lambda_3-\lambda_5) + q_1 = 0$$

$$(11.2) \qquad 1 + q_2 + \lambda_2 + \lambda_3 + \lambda_4 - \lambda_5 = 0$$

$$(11.3) \qquad \phi + \lambda_1 + \lambda_2 + \lambda_3 - \lambda_5 = 0$$

$$(11.4) \qquad \dot{q}_1 = (\rho + \delta)q_1 - (1 + \lambda_3)(1 - u_c)R_K - (1 + \lambda_3)u_c\delta p_I - \lambda_2 i$$

$$(11.5) \qquad \dot{q}_2 = (\rho + \psi)q_2 + (1 + \lambda_3)[(1 - u_c)r_b + \psi]$$

$$\lim_{t\to\infty} q_1 \geqq 0 \qquad \lim_{t\to\infty} q_1 K = 0 \qquad \lim_{t\to\infty} q_2 \leqq 0 \qquad \lim_{t\to\infty} q_2 B = 0.$$

There are also the relevant equations associated with the multipliers and $\dot{K}$, $\dot{B}$, as well as the Legendre-Clebsch conditions which state that the matrix, comprised of the derivatives of (11.1) to (11.3) with respect to I, b and s, is negative definite.

The optimality conditions illustrate that the interrelationship between the physical and financial decisions arises from the constraints and not from the existence of taxes. The tax rates affect the magnitude of the variables,but do not create an interdependence between the level and financing of investment.

3. Corporate Financial Policy

The determination of corporate financial policy centres around Equations (11.2) and (11.3), and it is governed by the value of $1 + q_2$. Whether internal or external funds (and which type of external funds) are used to finance physical investment depends on the value of the costate variable attached to the debt accumulation equation.

The meaning of q_2 can be discerned from the situation with $\lambda_3=0$ (positive dividends) and when ρ, u_c and r_b are time independent. In this case $\dot{q}_2=0$ $t \geqq 0$ and $\rho = (1-u_p)r/(1-u_g)$. Therefore, from (11.5)

$$q_2[(1-u_p)r/(1-u_g) + \psi] = -[(1-u_c)r_b + \psi].$$

The post tax marginal cost of corporate debt to the shareholders is $(1-u_c)r_b + \psi$. If the shareholders borrow the equivalent amount (by going short in the alternative asset) then the post tax marginal cost of personal debt is $(1-u_p)r/(1-u_g)$. Hence q_2 represents the difference between the post tax marginal costs of personal and corporate debt:

If $q_2 \gtreqless -1$ then the post tax marginal cost of personal debt $\gtreqless$ the post tax marginal cost of corporate debt.

The financing decision is also dependent on the fact that the capital gains tax rate is less than the tax rate associated with dividend income. This is reflected by $\phi = 1-1/a = (u_g-u_p)/(1-u_p)<0$. The marginal cost savings of either personal or corporate debt or corporate to personal debt must be compared to the personal marginal tax savings of capital gains over interest and dividend income. The tax deductibility of interest payments does not by itself create an advantage for bonds over equity financing, because corporate debt is a substitute for personal debt. However, the two types of debt are not perfect substitutes, because dividends cannot be paid out of the funds from new share and bond issues ($E \geqq 0$) and the regulatory authorities limit the amount of retentions ($E \leqq \ell p_I K$).

3.1 Financing Without Regulation

Let us begin the analysis of the financial policies by assuming that regulation does not exist or is ineffective, so that $\lambda_2=0$. First suppose that $1+q_2<0$. This yields the result (see the appendix) that $b+s=\max.(0, p_I I-F)$.

This permits us to characterize the financing decisions in terms of $1 + q_2$ to ϕ and $p_I I$ to F. In setting out the distinct cases recall that $E \geqq 0$, $D \geqq 0$, therefore $F \geqq 0$ and we shall only be concerned with the economically reasonable context of $I > 0$. The latter is not a limitation because reversible investment is not feasible and in the stationary state $I > 0$.

The different financing characterizations for $\lambda_2 = 0$ are depicted in Table 1. The derivation of this table is quite straightforward.

In the first case $\phi < 1 + q_2 < 0$ and $F < p_I I$. Due to Lemma 1, $b + s = p_I I - F$. If $F = 0$ then $E = D = 0$. If $F > 0$ with $b + s - p_I I = - F < 0$ and since $b + s - p_I I = -E$ then $E = F > 0$ and $D = 0$. From (11.2) and (11.3), $1 + q_2 - \phi = \lambda_1 - \lambda_4 > 0$, which means that $b > 0 (\lambda_4=0)$ and $s = 0 (\lambda_1>0)$ is the only combination feasible with $0 \leqq F < p_I I$. These results follow because the internal flow of funds is less than the value of physical investment, which creates a need for external financing sources. Debt is the external source since the net post tax marginal cost saving from personal debt is less than the tax savings from capital gains. Thus dividends are not paid out and any financing above retentions is derived from corporate bonds.

Table 1

Financing Characterization Without Regulation

	$1 + q_2 < 0$					$1 + q_2 \geqq 0$
	$F < p_I I$			$F > p_I I$	$F = p_I I$	
	$1 + q_2 > \phi$	$1 + q_2 < \phi$	$1 + q_2 = \phi$			
b	+	0	+	0	0	+
s	0	+	+	0	0	0
E	$F \geqq 0$	$F \geqq 0$	$F \geqq 0$	$0<E<F$	$F \geqq 0$	0
D	0	0	0	$D=F-E>0$	0	$F \geqq 0$
	(1)	(2)	(3)	(4)	(5)	(6)

Cases 2 and 3 are established in a similar fashion to case 1. When $1 + q_2 < \phi$ and $F < p_I I$, the tax saving from capital gains is less than the net marginal cost saving from personal debt. The firm issues new shares to obtain the external funds needed above retentions. If $1 + q_2 = \phi$ and $F < p_I I$, the tax saving from capital gains is less than the net marginal cost saving from personal debt. The firm issues new shares to obtain the external funds needed above retentions. If $1 + q_2 = \phi$ and $F < p_I I$, the shareholders are indifferent between debt and shares. The last two situations for $1 + q_2 < 0$ (that is,for $F \geqq p_I I$) are defined when internal funds exceed the value of physical investment. There is no need for new debt or shares, the firm pays dividends (for $F > p_I I$) and adds to accumulated earnings.

The number of cases diminish when $1 + q_2 \geqq 0$, because now $1 + q_2 > \phi$. In this context the post tax marginal cost of corporate debt is not greater than the post tax marginal cost of personal debt. As a consequence, corporate debt is the preferred financing instrument and all internal funds, if there are any, are paid out as dividends.

3.2 Financing With Regulation

Binding regulation (with $\lambda_2 > 0$) means that retentions have been effectively limited to $E = \mathit{i} p_I K$. Clearly if $\mathit{i} p_I K = F$ ($\mathit{i} p_I K > \dot{F}$ is impossible) then $E = F$ and as can be observed from Table 1, regulation is not an important restriction. Thus we reasonably assume that $\mathit{i} p_I K < F$, and so with the presence of a regulatory constraint, $E < F$ and $D = F - E > 0$. When authorities limit retentions per value of physical capital, this ceiling itself creates a flow of dividends to the shareholders.

There is another difference in the financing patterns which occurs with regulation. Solutions with $1 + q_2 \geqq 0$ without regulation imply that there are no retentions. However with regulation, $E = i p_I K > 0$, it is then impossible for $1 + q_2 \geqq 0$. Effective regulation creates a situation where the post tax marginal cost of personal debt is less than the post tax marginal cost of corporate debt. The intuition behind this conclusion follows by separately considering the situations $1 + q_2 \geqq 0$ and $1 + q_2 < 0$. First suppose that $1 + q_2 \geqq 0$ in the absence of regulation. In this case the dividend payout rate is unity ($D/F=1$). Once regulation becomes effective, the firm is forced to decrease its payout rate since retentions are positive. We find that the post tax marginal cost of personal debt falls relative to the post tax marginal cost of corporate debt. Corporate debt becomes more expensive and the firm cannot afford a payout rate of unity.

Next suppose $1 + q_2 < 0$ is relevant. We want to compare the values of q_2 with and without regulation. In order to do this (and subsequently to describe a stationary state) we assume that ρ, u_c and r_b are time independent. This implies, from Equation (11.5), that $\dot{q}_2=0$ for all $t \geqq 0$. For $\lambda_2=0$, $1 + q_2 < 0$ and $F \leqq p_I I$, from Table 1 $D=0$ and $\lambda_3 > 0$, so

$$q_2^m = -(1 + \lambda_3)[(1-u_c)r_b + \psi]/(\rho + \psi)$$

where q_2^m is the value of q_2 without regulation. If $\lambda_2 > 0$ so $1 + q_2 < 0$

and with $F \leqq p_I I$, then $D > 0$, $\lambda_3 = 0$, and

$$q_2^r = - [(1-u_c)r_b + \psi] / (\rho + \psi)$$

where q_2^r is the solution with effective regulation. Obviously, $q_2^m < q_2^r < 0$. With the move to regulation, q_2 increases thereby causing a relative decrease in the post tax marginal cost of corporate debt. Regulation renders bond (and share) markets relatively more attractive, because of the limitation on the firm's ability to use internal funds to finance investment.

Table 2 summarizes the financing policies for a regulated firm. Our model exhibits both a relatively stable dividend payout rate and a frequent use of financial capital markets for regulated as opposed to unregulated firms.[8]

Table 2

Financing Characterization With Regulation

	$1 + q_2 < 0$			
	$F \leqq p_I I$			$F > p_I I$
	$1 + q_2 > \phi$	$1 + q_2 < \phi$	$1 + q_2 = \phi$	
b	+	0	+	0
s	0	+	+	0
E	$i p_I K > 0$	$i p_I K > 0$	$i p_I K > 0$	$i p_I K > 0$
D	$F - i p_I K > 0$	$F - i p_I K > 0$	$F - i p_I K > 0$	$F - i p_I K > 0$
	(1)	(2)	(3)	(4)

4. Short Run Equilibrium

Given values for K, B, q_1, and q_2 we investigate the nature of physical investment, new debt and share issues.

4.1 No Regulation in the Short Run

The first case defined by $b > 0$, $s = 0$, $E = F \geq 0$ and $D = 0$ means that (with $F > 0$) Equations (11.2) and (11.3) become $1 + q_2 + \lambda_3 = 0$, $\phi + \lambda_1 + \lambda_3 = 0$. Thus from (11.1) and with D = 0, I and b are determined from

$$q_2 (1 - u_c)A' + p_I q_2 + q_1 = 0$$

$$(1-u_c)[K(K,t) - A(I) - r_b B] + u_c \delta p_I K + b - \psi B - p_I I = 0.$$

Therefore, $I = I^{m1}(q_1,q_2)$ with $\frac{\partial I}{\partial q_1} = -1/q_2(1-u_c)A'' > 0$ and $\frac{\partial I}{\partial q_2} = -[1(1-u_c) A' + p_I]/q_2(1-u_c)A'' > 0$. When the demand price of investment and when the post tax marginal cost of corporate debt increases relative to that for personal debt, the firm increases its physical investment. For new debt, $b = B^{m2}(q_1,q_2, K ,B)$ with $\frac{\partial b}{\partial q_i} = - [(1-u_c)A' + p_I] \frac{\partial I}{\partial q_i} > 0$ $i=1,2$. Increases in investment are accompanied by corresponding increases in debt issues.

In case 2 with $b = 0$ $s > 0$ $E = F > 0$ and $D = 0$, investment and share issues are determined from

$$(\phi - 1)(1 - u_c)A' + p_I(\phi - 1) + q_1 = 0$$

$$(1 - u_c)[K(K,t) - A(I) - r_b B] + u_c \delta p_I I - \psi B + s - p_I I = 0.$$

We find that $I = I^{m2}(q_1)$ with $\frac{dI}{dq_1} = 1/(1 - \phi)(1 - u_c)A'' > 0$, while

$s = S^{m2}(q_1,K,B)$ and $\frac{\partial s}{\partial q_1} = [1 - u_c)A' = p_I] \frac{dI}{dq_1} > 0$. Investment and new share issues do not respond to changes in q_2, and both increase as the demand price for investment rises. We get the same results for case 3 (as for case 2) with respect to investment. However, in this case we are unable to determine s and b separately, we can only solve for s + b; the value of new financing is determinate while the composition is irrelevant.

In case 4, $b = s = 0$, $E > 0$, $D > 0$, and the equation determining investment (from (11.1)) is

$$-(1-u_c)A' - p_I + q_1 = 0.$$

Hence $I = I^{m4}(q_1)$ and $\frac{dI}{dq_1} = 1/(1-u_c)A'' > 0$

Case 5 is on the surface similar to case (4), but since $D = 0$ investment is determined from the dividend equation. Consequently, $I = I^{m5}(K,B)$; investment is independent of its demand price.

Finally for case 6 with $b > 0 \quad s = 0 \quad E = 0 \quad D = F > 0$, we have $p_I I = b$ and investment is determined from,

$$-(1 - u_c)A' + p_I q_2 + q_1 = 0.$$

Thus $I = I^{m6}(q_1,q_2)$ with $\frac{\partial I}{\partial q_1} = 1/(1-u_c)A > 0$ and $\frac{\partial I}{\partial q_2} = p_I/(1-u_c)A''>0$ while $b = B^{m6}(q_1,q_2)$ and $\frac{\partial b}{\partial q_i} = p_I \frac{\partial I}{\partial q_i} > 0 \quad i=1, 2.$

The importance of these results is two-fold. First, they point out how the short run investment demand function differs according to the financing patterns. Table 3 summarizes these distinct demands. Second, the short run results are needed to analyse the dynamics and stationary state behavior of the corporation. Before proceeding to the dynamics, we discuss the short run equilibrium in the presence of effective regulation.

4.2 Regulation in the Short Run

Case 1 defined for $b > 0$, $s = 0$ $E = i p_I K$ and $D > 0$ leads to investment and new debt being determined from

$$-(1-u_c)A' + p_I q_2 + q_1 = 0$$

$$p_I I - b - i P_I K = 0.$$

Therefore $I = I^{r1}(q_1, q_2)$ with $\frac{\partial I}{\partial q_1} = 1/(1-u_c)A''>0$ and $\frac{\partial I}{\partial q_2} = p_I/(1-u_c)A'' > 0$. Interestingly, in this context the investment demand function is identical to that found for case 6, in the absence of regulation. This occurs because in both situations debt is the financing instrument and dividends are paid out to the shareholders. For debt, $b = B^{r1}(q_1, q_2, K)$ with $\frac{\partial b}{\partial q_i} = p_I \frac{\partial \dot{I}}{\partial q_i} > 0$ $i = 1, 2$. The ceiling on retentions causes the change in bonds to be equal to the value of the change in investment.

The next solution where $b = 0$, $s > 0$ $E = i p_I K$ and $D > 0$ leads to

$$-(1-u_c)A' + p_I(\phi - 1) + q_1 = 0$$

with $I = I^{r2}(q_1)$, $\frac{dI}{dq_1} = 1/(1-u_c)A'' > 0$ and $s = S^{r2}(q_1, K)$, $\frac{\partial s}{\partial q_1} = p_I \frac{\partial \dot{I}}{\partial q_1} > 0$. We can observe that $I^{r2}(q_1) = I^{m4}(q_1)$. Case 4 is the other case (besides 6)

Table 3

Short Run Investment Demand Functions

	No Regulation		Regulation
Case		Case	
1.	$I^{m1}(q_1,q_2)$	1.	$I^{m6}(q_1,q_2)$
2.	$I^{m2}(q_1)$	2.	$I^{m4}(q_1)$
3.	$I^{m2}(q_1)$	3.	$I^{m4}(q_1)$
4.	$I^{m4}(q_1)$	4.	iK
5.	$I^{m5}(K,B)$		
6.	$I^{m6}(q_1,q_2)$		

where dividends are paid in the absence of regulation. Thus in conjunction with an identical financing pattern, case 2 with regulation, corresponds in terms of investment, to case 4 without regulation. Case 3 is similar to case 2 with respect to investment, although we are only able to solve for $b + s = S^{r2}(q_1,K)$.

In the last case, because retentions are $ip_I K$ and there is no external financing, then $I = iK$. The investment-capital ratio is determined by the regulator and it is fixed for all $t \geqq 0$.

Because of the correspondence between the short run investment demand functions, it is worthwhile comparing the investment response to changes in its demand price for similar financing patterns and across regulatory regions. For the first three cases in Table 3, we find with $1 + q_2 < 0$ and $1 - \phi > 1$ that $\frac{\partial I^{m4}}{\partial q_1} = \frac{\partial I^{m6}}{\partial q_1} > \frac{\partial I^{mi}}{\partial q_1} > 0$ $i = 1,2$. Hence investment, under regulation responds more to change in its demand price. The reason is that given q_2, with binding regulation, the ceiling on retentions limits the ability of the corporation to undertake investment. The firm desires more investment per marginal dollar spent relative to the unregulated situation, because it cannot redirect internal funds for investment.

5. Dynamics and the Stationary State

The optimal path to be followed by the firm depends on whether or not regulation is binding and which financing pattern arises. The relevant differential equations describing the dynamics are:

$$(12.1) \qquad \dot{K} = \mathcal{I}(q_1, q_2, K, B) - \delta K$$

$$(12.2) \qquad \dot{B} = \mathcal{B}(q_1, q_2, K, B) - \psi B$$

$$(12.3) \qquad \dot{q}_1 = (\rho + \delta)q_1 - (1 + \lambda_3)[(1 - u_c)R_K + u_c \delta p_I] - \lambda_2 i \, p_I$$

$$(12.4) \qquad \dot{q}_2 = (\rho + \psi)q_2 + (1 + \lambda_3)[(1 - u_c)r_b + \psi].$$

In order to be able to depict the different stationary states we must assume that $R(K,t)$ converges to a value which is time independent. That is after some $t \geq 0$ technological change does not affect the indirect variable profits function. We also assume that p_I is stationary.[9]

5.1 No Regulation in the Long Run

When regulation is not effective (so that $\lambda_2 = 0$), there are six possible cases to consider. These cases can be shown in a diagram in $(K, -q_2)$ space.

Cases 1 and 3 cannot be stationary states because from Equation (12.4), $\dot{q}_2 = [\rho - (1-u_c)r_b]q_2$. Now with $\rho \neq (1-u_c)r_b$ which is implied by cases 2, 4, 5 and 6 and since the values of ρ, u_c, r_b are stationary, $\dot{q}_2 = 0$ for all $t \geq 0$ with $q_2 = 0$.[10] But in cases 1 and 3, $1 + q_2 < 0$,

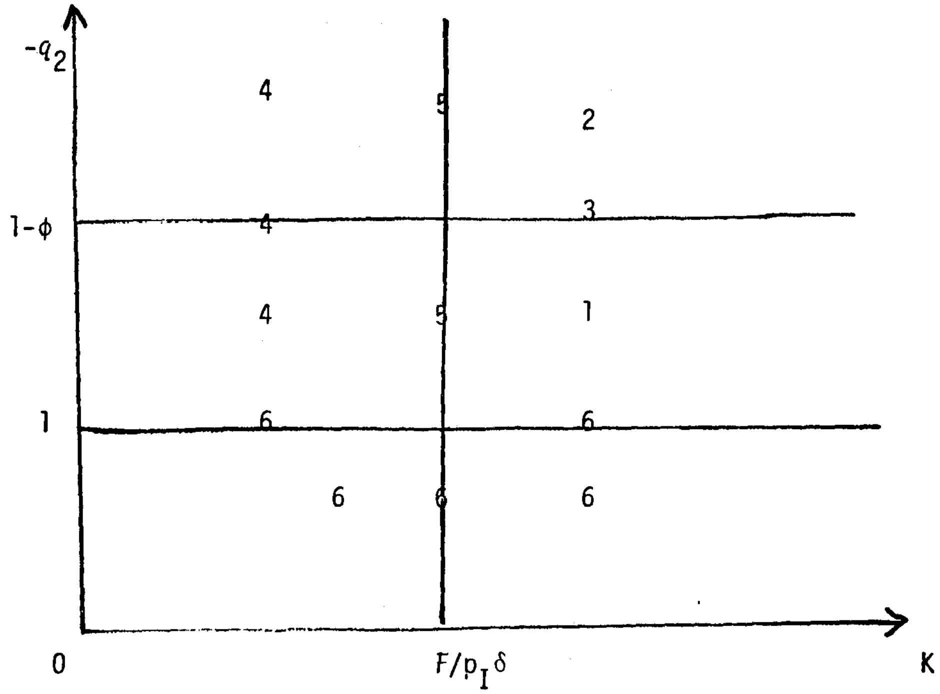


Figure 1

Financing Patterns Without Regulation

and so we get a contradiction. Hence, because the stationary state cannot occur in 1 or 3, we consider the dynamics for cases 2, 4, 5 and 6.

For the second case, Equation set (12) becomes,

$$\dot{K} = I^{m2}(q_1) - \delta K$$

$$\dot{B} = -\psi B$$

$$\dot{q}_1 = (\rho + \delta)q_1 + (\phi - 1)(1 - u_c) R_K + (\phi - 1) u_c \delta p_I$$

$$\dot{q}_2 = (\rho + \psi)q_2 + (1 - \phi)[(1 - u_c) r_b + \psi].$$

Clearly $\dot{q}_2 = 0$ for all $t \geq 0$ and $q_2^{m2} = -(1 - \phi)[(1-u_c)r_b + \psi]/(\rho + \psi)$.

Moreover because $b = 0$ and with $\dot{B} = 0$ then all debt is retired in this stationary state. From the $\dot{K}$ and $\dot{q}_1$ equations,

$$\frac{\partial \dot{K}}{\partial q_1} = \frac{\partial I^{m2}}{\partial q_1} > 0, \quad \frac{\partial \dot{K}}{\partial K} = - \quad \delta < 0$$

$$\frac{\partial \dot{q}_1}{\partial q_1} = \rho + \delta > 0 \quad , \frac{\partial \dot{q}_1}{\partial K} = (\phi - 1)(1 - u_c)R_{KK} > 0.$$

Thus the stationary state is unique and a saddle point, if it occurs in case 2, $0<K^{m2}< \infty$ while $0 < q_1^{m2} = (1 - \phi)[(1-u_c)R_K + u_c\delta p_I]/(\rho + \delta)$.

In a similar manner we can establish that if the stationary state exists in case 4, 5, and 6, then it is unique and a saddle point, with $0 > q_2^{m4} = [(1-u_c)r_b + \psi]/(\rho + \psi) = q_2^{m6} > q_2^{m2}$. In all stationary states, except for case 6, the firm retires all debt, and does not issue any new shares.

We can compare K^{m6} to K^{m2} and K^{m4} in order to derive the comparisons, utilize Equations (11.1) - (11.3), (12.3) and (12.4), to find the marginal revenue product of capital and the associated user cost. For cases 2 and 4, that is for $1 + q_2 < 0$

$$(13) \quad [R_K - (\rho + \delta)A'] = p_I(\rho + \delta(1 - u_c))/(1 - u_c)$$

where the right side of (13) is the user cost of capital. Notice that if the personal and corporate income tax rates are equal and there is no capital gains tax, then $\rho = r(1 - u_c)$. The user cost becomes $p_I(r + \delta)$; the income tax rate is neutral. For case 6 $(1 + q_2 \geqq 0)$,

$$(14) \quad [R_K - (\rho + \delta)A'] = \frac{(\rho+\delta)}{(\rho+\psi)} p_I[r_b + \frac{1}{(1-u_c)} (\psi - u_c\delta\frac{(\rho + \psi)}{(\rho + \delta)})]$$

where the user cost is the right side of (14). The reason for the complication is because physical capital is depreciated at a rate which is not necessarily the same as debt is retired. If $\delta = \psi$ then the user cost is $p_I(r_b + \delta)$. Now taxes are neutral even when $u_p \neq u_c$ and $u_g > 0$, because the relevant rate of return is the corporate bond rate and not the shareholders discount rate. The key distinction in the user cost of capital is whether equity (retentions or shares, when $1 + q_2 < 0$) or debt $(1 + q_2 \geqq 0)$ is used to finance investment. The financing pattern does matter.

If $\delta = \psi$ and let the right side of (13) be w_K^{m2} and the right side of (14) be w_K^{m6}, then $w_K^{m6} - w_K^{m2} = p_I(r_b - \rho)/(1 - u_c) < 0$. Thus

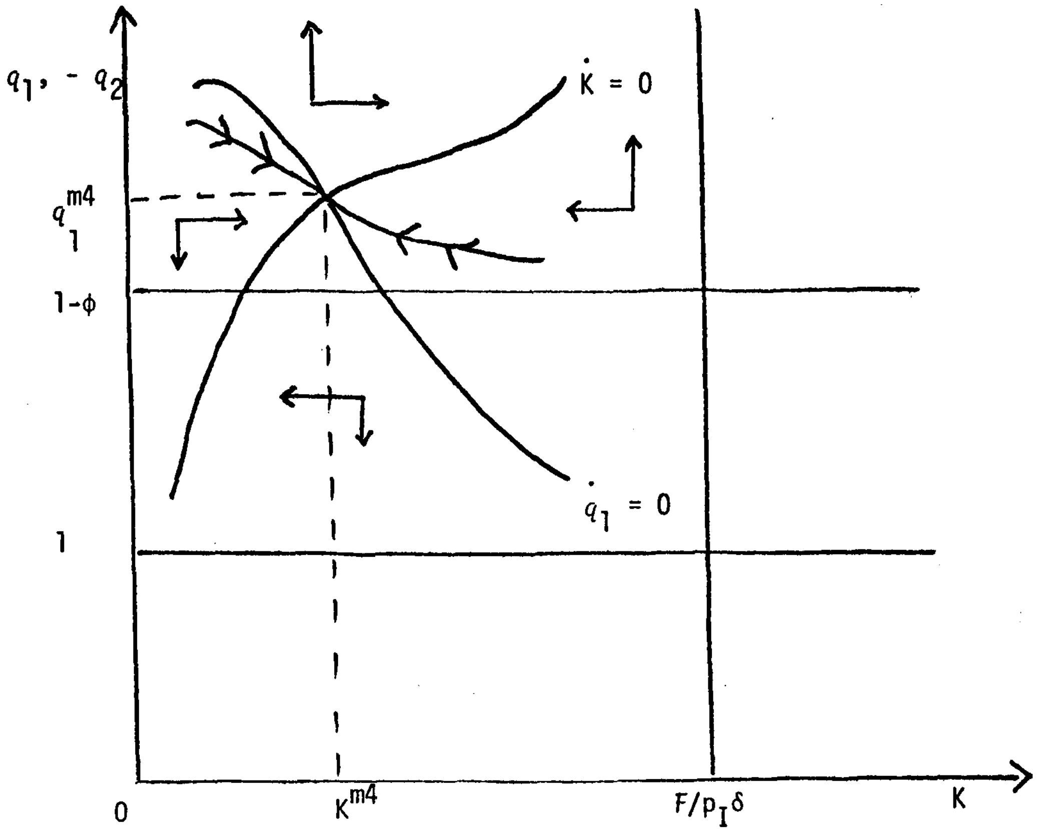


Figure 2

Stationary State in Case 4 Without Regulation when capital depreciates at the same (or smaller) rate than bonds retire,then $K^{m6} > K^{m2} = K^{m4}$. Figure 2 illustrates the stationary state which can arise in case 4.

5.2 Regulation in the Long Run

In the presence of binding regulation $1 + q_2 \geqq 0$ is not feasible. Thus the possible financing patterns in $(K,-q_2)$ depicted by Figure 3 show $1 + q_2 < 0$. There are four possible patterns as described in Table 2.

Unlike the unregulated firm, stationary states can exist in cases 1 and 3, because retentions are less than the internal flows of funds,so there are positive dividends. For the first case, $q_2^{r1} = q_2^{m4}$, (12.1) and (12.3) become

$$\dot{K} = I^{m6}(q_1, -[(1 - u_c)r_b + \psi]/(p + \psi)) - \delta K$$

$$\dot{q}_1 = (\rho + \delta)q_1 - (1 - u_c)R_k - u_c\delta p_I + i p_I - i p_I[(1-u_c)r_b + \psi]/(\rho + \psi).$$

Clearly in (K,q_1) space, $\dot{K}=0$ is positively sloped and $\dot{q}_1 = 0$ is negatively. The unique stationary state is a saddle point, at $0<K^{r1}<\infty, q_1^{r1} > 0$.

There is a further interesting result to mention concerning the long run equilibrium. Since $I = \delta K$, $b = \psi B$ and $E = i p_I K$, then,

$$p_I(\delta - i)K = \psi B > 0$$

for case 1. Thus for the long run equilibrium to exist i must be set less than δ , by the regulatory authority. The retention-asset ratio must be less than the stationary investment to capital ratio.

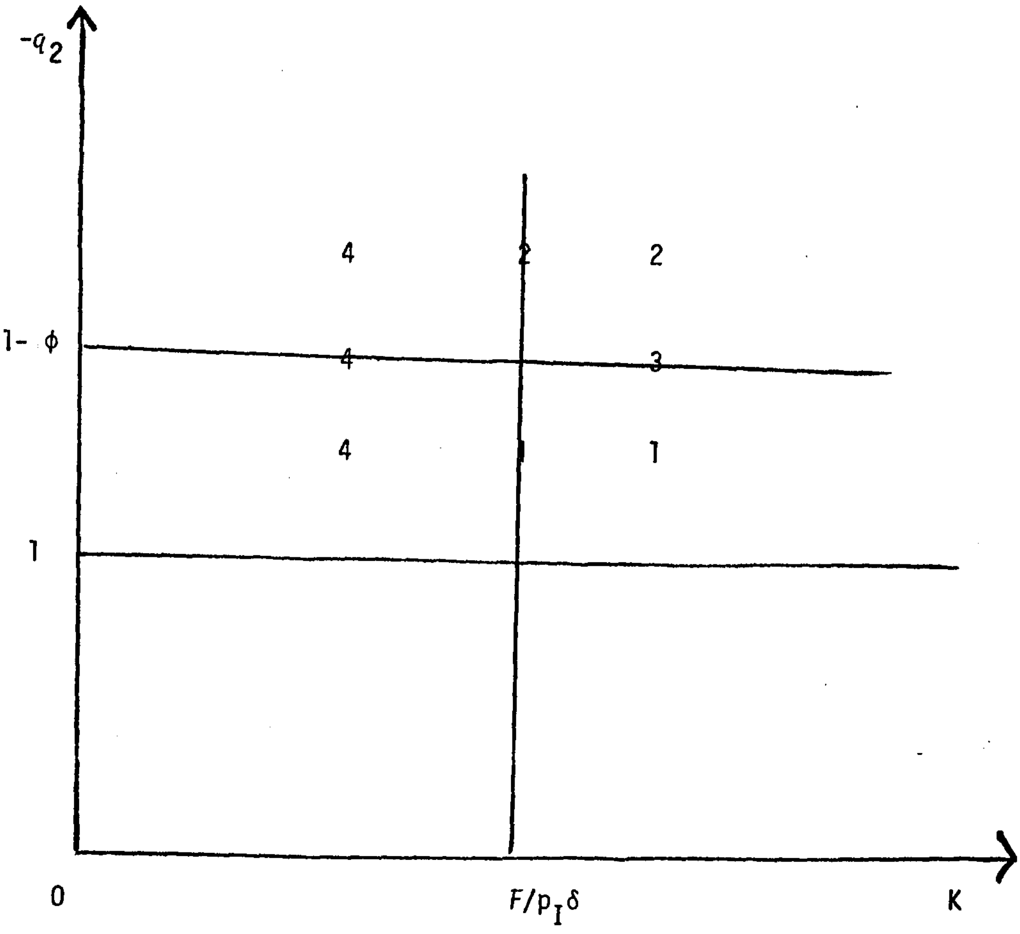


Figure 3

Financing Patterns with Regulation

We can in an identical manner describe the dynamics and stationary state for cases 2-4. We find that q_2^{m4} is the value of q_2 in all the stationary states, while all debt is retired in cases 2-4.[11] Figure 4 illustrates the stationary solution for case 2.

By inspecting Figures 1 and 3, it is obvious that it is possible for the physical capital stock to be smaller under regulation in the stationary state. To be more precise let us develop the user cost of capital for the first two cases under regulation.[12] For case 1, using Equations (11.1) - (11.3), (12.3) and (12.4) we get,

$$(15) \quad R_K - (\rho + \delta)A' = w_K^{m6} + (1 + q_2^{m4})i p_I = w_K^{r1}.$$

Since $1 + q_2^{m4} < 0$ then $w_K^{r1} < w_K^{m6}$ and consequently $K^{r1} > K^{m6}$. In addition if $\delta \leqq \psi$ (for example with one period bonds) then we saw that $K^{m6} > K^{m4} = K^{m2}$. Therefore the capital stock is the largest when the stationary state (with regulation) occurs for case 1. Because debt financing is used, the interest deductibility provision permits the firm to meet the regulatory constraint and still expand its capital stock.

The solution is rather different when $1 + q_2 < \phi$ and equity is used to finance investment. The user cost of capital when new shares are used to finance investment is

$$(16) \quad R_K - (\rho + \delta)A' = w_K^{m2} - \phi p_I(\rho + \delta - i)/(1 - u_c) = w_K^{r2} .$$

Since $\delta - i > 0$ for the solution to exist and $\phi < 0$, then $w_K^{r2} > w_K^{m2}$.

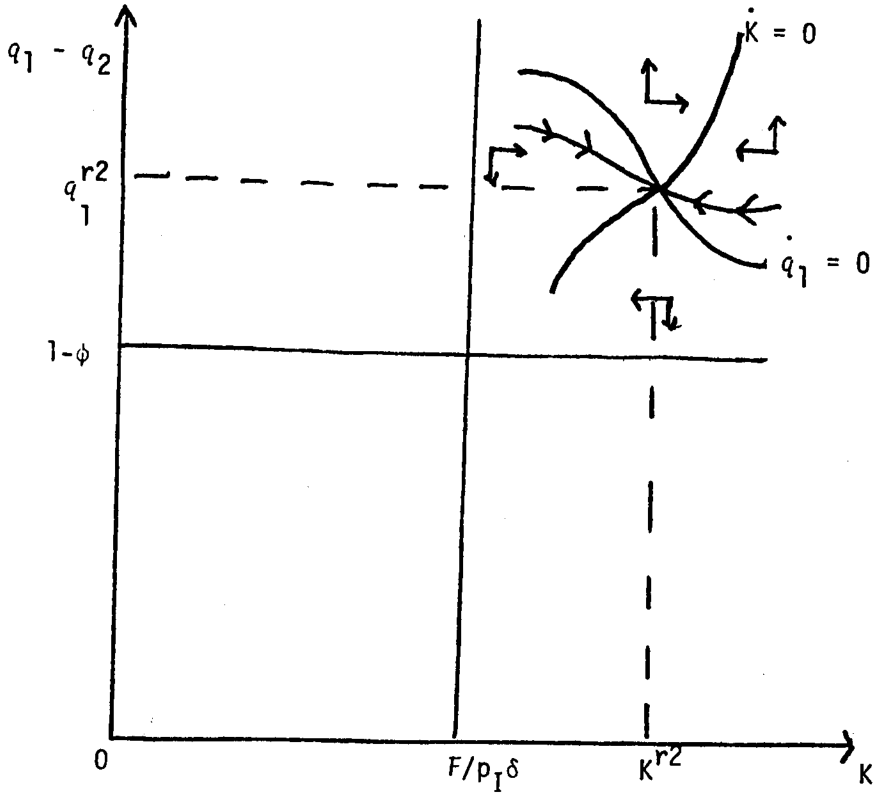


Figure 4

Stationary State in Case 2 With Regulation

Therefore $K^{r2} < K^{m2} = K^{m4}$. The capital stock under regulation is smaller. Intuitively, with internal funds and new shares as the financing instruments, the firm devotes part of its funds to retiring all of its debt, while simultaneously facing an upper bound on its retentions. Therefore in order to satisfy the regulator, it must have a relatively smaller capital stock. Moreover, if $\delta \leqq \psi$ then $K^{r1} > K^{m6} = K^{m2} > K^{r2}$, the capital stocks in the regulated solutions form the upper and lower bounds to the stocks which can appear in the unregulated stationary states.

6. Conclusion

The results of this paper point out the contrasts of the financial and investment policies of regulated firms to their unregulated counterparts. We established that unregulated corporations tend to finance investment through internal funds, while regulated firms utilize financial markets to a greater degree. The upper bound on retentions forces the regulated corporation to direct funds to dividends, thereby affecting the investment and physical capital decisions. We noted that if the regulated firm finds it cheaper to finance through the bond market, then we observe a relatively larger capital stock. However, a regulated firm which tends to issue shares will have a smaller stock. Not only is this conclusion in contrast to the static model, but we are able to derive a matching of the financial policy to the size of the capital stock.

Although we have come some distance, much work remains to be done, especially with regard to short run deviations from the regulatory constraint, because of asymmetric information on the part of the firm and regulator. The problem of financing in the presence of stochastic regulatory review is complex and remains unsolved.

FOOTNOTES

1. For notational convenience t is now a subscript, when it signifies the time dependence of a variable.

2. We are assuming that borrowing and lending rates on corporate debt are identical.

3. A dot over the variable signifies the derivative with respect to time.

4. For an analysis of regulatory lag see the paper by Klevorick [10].

5. Capital gains are taxed on an accrual basis rather than when they are realized, and the tax rates are independent of their respective bases. This is the approach followed by Stiglitz [14] and King [9].

6. The market value of the shares is defined over an infinite horizon. A finite horizon raises questions of the length of the horizon, the terminal value of physical capital, and the disposition of shares and bonds. However, with a complete set of markets for capital (physical and financial) our results can be specialized to the finite horizon case.

7. We drop the time notation, unless it is necessary for clarification.

8. Auerback [2] notes the infrequent issue of shares by unregulated firms.

9. In order to guarantee that $0 < K < \infty$ in the stationary state we assume that $R_K > (\rho + \delta)q_1/(1 - u_c)(1 + \lambda_3) - u_c\delta p_I/(1 - u_c)$ for $K = 0$ and $R_K < (\rho + \delta)q_1/(1 - u_c)(1 + \lambda_3) - u_c\delta p_I/(1 - u_c)$ for $K = \infty$.

10. $\rho \neq (1 - u_c)r_b$ is needed in order for individuals not to be indifferent between shares and corporate bonds.

11. In order for a stationary state to exist in cases 1-3 $\delta > i$. However, for case 4 to be consistent with a stationary state $\delta = i$ and so $\dot{K} = 0$ for all $t \geq 0$ and thus $I = \delta K$ for all $t \geq 0$. In this situation investment immediately adjusts to its stationary value. This is unlike the unregulated case where investment can be characterized by a flexible accelerator (see Treadway [15]).

12. Case 3 is identical to case 2 and case 4 cannot exist if $\delta \neq i$.

Appendix

Lemma 1. If the firm maximizes V_0 according to Equation set (11) with $\dot{K} = I - \delta K$, $K_0>0$, $\dot{B} = b - \psi B, B_0>0$, and $b\geq 0$, $D\geq 0$, $s\geq 0$, $0\leq E< p_I K$ and $I>0$ then $b + s = \max. (0, p_I I - F)$.

Proof. Suppose $b + s \neq \max.(0, p_I I - F)$. In addition, first assume $b + s \neq \max. (0, p_I I - F) = 0$. In this case either $b + s < 0$ or $b + s >0 > p_I I - F$. Clearly $b + s < 0$ is impossible. Also $b + s > 0 > p_I I - F$ implies that $b + s > b + s - D$, since $F= E + D$ and then $D > 0$ $(\lambda_3 = 0)$. With $\lambda_3 = 0$ Equations (11.2) and (11.3) become

$$1 + q_2 + \lambda_4 = \lambda_5$$

$$\phi + \lambda_1 = \lambda_5.$$

If $\lambda_5> 0$ then we see that if λ_1 or λ_4 equal 0 then $1 + q_2 = \lambda_2$ or $\phi = \lambda_2$ which cannot be true. Moreover $\lambda_1 > 0$, $\lambda_4 > 0$ means $b = s = 0$ which is not possible. Hence $\lambda_5 = 0$ (and $E > 0$). When $\lambda_5 = 0$ we get contradictions because if $\lambda_4 > 0$ then $\lambda_1 = 0$ and $\phi = 0$ which is not true. Similarly if $\lambda_4 = 0$ then $1 + q_2 = 0$, which is not true.

For the second case we assume $b + s \neq \max (0, p_I - F) = p_I I - F$. Here either $b + s < p_I - F$ or $b + s > p_I I - F > 0$. If $b + s < p_I I - F$ then it is implied that $E > E + D$ which is impossible since $D \geq 0$. If $b + s > p_I I - F > 0$ then $E + D > E$ and so $D > 0(\lambda_3 = 0)$. We then have $1 + q_2 + \lambda_4 = \lambda_5$, $\phi + \lambda_1 = \lambda_5$ which we previously showed to lead to contradictions. Thus $b + s = \max.(0, p_I I - F)$.||

A Cost-based Tariff Policy, Integrated Network Use and Network Competition*

by Jürgen Müller

I - Introduction

Linking financial and economic analysis immediately leads us on to the issue of tariffs, on which there exists vast literature[1]. It is not the purpose of this paper to survey this literature, but to depart from it by advocating cost based tariffs. Most of the literature has looked at telephone tariffs as an issue of second best pricing in the Ramsey-Boiteux tradition. Our reasons for cost based tariffs are :

- sustainable under network competition
- probably only small welfare losses as compared to second best tariffs,
- ease of implementation,
- control of political pressures on the tariffs,
- clear signaling effect with respect to long term investment policy
- long term dynamic benefits.

This proposition for a cost oriented tarif is not made in isolation, put together with the recommendation for network competition, either in the form of service or of facility competition.

If one wants to move in this direction, some cost oriented tariffs are a necessary consequence. Even if one moves only towards partial work competition on the basis of a shared use and resale policy (i.e. service competition) this proposition holds. But as we show below, a cost-oriented tariff policy is not only corollarly of a move towards network competition, but offers in itself some interesting properties. We will develop them below.

Such a policy only makes sense, of course, if the attributable costs are a significant part of total costs. Even though some experts may dispute this point, I believe this generally to be the case. But the issue is not so much an empirical question, but a conceptual question. We will show that with an "ideal" accounting system, the proportion of non-attributable cost is equivalent to the degree of economies of scale within the system. Large economies of scale are associated with a smaller proportions of attributable costs. Already there is ample evidence to suggest, that given the size of the telecommunications system in most developed nations, the remaining economies of scale are not very large. This suggests, that, as proposed above, attributable costs are a significant part of total costs. It is therefore not only possible, but also economically desirable to move towards a tariff system in which the tarif structure is more or less a mirror image of attributable costs.

The details of this argument will be outlined in Section III of this paper. In the next section, we discuss the corollary of cost based tariffs, namely the arguments for network competition.

II- From Monopoly Networks to Competitive Offering

1) Competition as a Regula tory Tool

In the past and in most countries today, the telecommunications network is considered to be a natural monopoly and therefore protected from competition and entry through legal barriers. The service providers, who operate this legal monopoly are either privately regulated firms or, more commonly, publicly owned corporation. Empirical evidence suggests however, that both privately owned and regulated corporations and state owned enterprises have difficulty fulfilling their regulatory goals.[2] They therefore need to be supervised continuously. Competition as the regulatory tool is, at least this can be suggested form the experience in the US (and perhaps Canada as well) an effective policy tool in this respect[3].

...

For example, competition as a regulatory tool will help to avoid biased investment decisions, such as the Averch-Johnson[4] Effect. It should also change the often quite conservative investment and operating policy of state owned enterprises. For example, many of them tend to value the reliability of service and avoidance of interruptable service (poor quality would give them a poor public image) often higher than their own customers would. Network competition would allow such customers to choose their own level of quality, therefore acting both as a signal to the service providers and the regulators.

2) Competition as an Efficient Search Process

I argue for a very liberal network policy, because unrestricted user and producer freedom we will allow competition to function as an efficient search process[5]. This is especially important in such sectors where the search for the applications of new technology potential is important. Telecommunications is one of these sectors. By allowing entry into network competition, even if only by a shared use and resale policy, will stimulate search processes for new applications, in addition to leading towards a more efficient network utilization. Such private networks and also value added networks (VAN) not only change the behaviour of the service providers, but also act as a signal for new technology applications to the PTT. Because of their requirement to serve, they usually tend to be slow to innovate, unless the product has proven itself beyond doubt and a nationwide demand can be established. Competitive entrants do not have this obligation to serve; they therefore face a smaller risk in penetrating only some segments of the market and therefore act as important signal providers of the available technology and demand potential for

...

for the larger PTTs. The potential of the market to act as a successful search process is therebyconsiderably increased.

3) Departure from Uniform National Tariffs and Cream Skimming

One of the most important arguments against increased user liberalisation (including resale and shared use) is the issue of cream skimming. Users can transform certain particular services into others and to resell them once, thereby reducing the income available for the PTT. Even simple arbitrage has the same effect. In most case, such cream skimming will be a direct result of user liberalization, since the tariffs are not proportional to costs. This is normally the case with uniform national tariffs in the case of significant cost variations. But it is quite easy to eliminate such cream skimming, when network operators move towards cost-oriented tariffs. In this case, only such networks which are more efficient than the PTT or offer some enhanced service will be able to survive. In other works, those enterprises planning to survive on arbitrage alone will probably have a short lived existence.

4) Departure from Cross Subsidization

Insistence on uniform national tariffs in the face of significant costs variations also implies significant cross-subsidization. While in the past, such policies have been explicitly designed to increase network penetration, for example, by subsidizing private, price elastic household demand from the less price elastic business sector, the degree of network penetration is so high, that this argument, at least in the most developed countries, no longer carries much weight. Even those, who believes that cross-subsidization is an efficient redistributive tool, especially when others may be politically unfeasible[6], must see that the benefits of being able to cross subsidize between servies may not be very large and can easily be out-weight by their associated efficiency losses. These are the associated

...

inefficiencies in investment and expenditure on substitutes for those services, which are priced above costs and the reduced incentive to innovate for those services which are priced below cost. Further inefficiencies are caused by lobbying efforts, giving a strong preference for maintaining the status quo by those disadvantaged by technical changes(7). The threat of entry will not only reduce the available surplus for cross-subsidization and make prices more efficient, but will also reduce the potential of politicians to tinker with the system in a politically opportune way.

Even if contined cross-subsidization is desired, user's freedom may not have to be curtailed under a cost oriented tariff. As we shall show below, a cost oriented tariff is based on a proportional charge above the attributable costs in order to yield a final tariff. This proportional mark-up could also form the basis of resale policy. It would be based on a difference between the retailer's tariff and his payments for leased lines. The final user pays, in this way, the same contribution to the financing of the non-attributable costs as do other users. In other words, this procedure is equivalent to the value added tax but as we shall see below, it complicates the calculations somewhat.

Departure from Value of Service Based Tariffs

While many PTT Administrations have usually not pursued second best pricing in the Ramsey-Boiteux tradition, they often have pursued value of service pricing. While such tariffs take some cost considerations into account, they are usually more closely related to the value which the customer places on these services. For example, long distance and international calls are therefore prices very much above cost. This policy allowed the PTTs to cross subsidise local calls and access charges and thereby achieve a higher degree of network penetration than with purely cost based tariffs. But see a counter argument below).

To some extent, this pricing policy also resembles that of a discriminating ponopolist and, at least in its pricing structure (but not in its price level), second best pricing (under the assumption that the willingness to pay rises with the value which the user attachs to it, while the price elasticity decreases). This policy may have been relevant in order to achieve a high level of network penetration, but has, with respect to the current degree of network penetration in most developped countries outlived its usefulness. Its continued existence can only be achieved by further user restictions and entry barriers in order to prevent cream skimming from taking place. This goes, of course, against our proposition to increase user and producer freedom to improve the functioning of the market as a search process.

6) Departure form Multiple Service Networks

With the move towards fully integrated digital networks, the current boundaries between existing services became very blurred. While some of these boundaries may continue to be legally upheld in order to continue a certain amount of price discrimination associated with value of service pricing (and to reduce the price elasticity of demand by making it more difficult to use substitute services). Such restrictions again unduly limit user and producer freedom. This is specially important with respect to the emergence of many new telecommunication applications, in which decentralized decision making may expoit the available technology potential more fully. A move towards a cost based tariff is therefore a very attractive option, which will allow the market to develop the potential of a fully integrated, digital network. This may still lead to a number of seperate networks on the basis of a fully integrated digital network, but the relation between them is now only based on cost, not value of service criteria.

- A COST-ORIENTED TARIFF

1) Attributable Costs

I have argued above for a cost oriented tariff.The charges for each service should be proportional to those costs which can be attributed to that service. Such a proposal only makes sense if most costs can be attributed directly. If in the extreme, only 10 % of costs are attributed, such a policy no longer makes sense. We must therefore show that most costs are attributable, provided the accounting system can be fine tuned to that extend.

The basic argument for a significant attribution of costs rests on the following theoretical proposition. Take a cost function

$$(1) \quad K = K(X_1, X_2, \ldots. \quad X_n),$$

where K are total costs and X_1, X_2, ... X_n represents the levels of each of the output activities of the firm/plant in question. For the purpose of simplifying the illustration, we assume that K is contineously differenciable for all the outputs. We can then define the marginal cost as the partial derivatives with respect to each output, in this case K_1, K_2,... K_n. We now define the attributable costs with respect to output i through $x_i K_i$. Total attibutable costs, according to this definition, are then

$$(2) \quad Z = \sum_{i=1}^{n} x_i K_i$$

It is now easy to show that the proposition of attributable costs to total costs is equal to r, the ratio of relevent marginal costs to average costs of the firm/plant in question, i.e.

$$(3) \quad \frac{Z}{K} = r$$

But 1/r is only a measure for scale effects within the system.[8] If one trusts the empirical evidence currently available, then the scale effects of todays telecommunication systems are not very large, given their development and network penetration. In this case, 1/r is only marginally larger than Z and r pretty close to 1. A large proportion of costs in this sense is therefore attributable.

Some may question this definition of attributable costs. But it does make economic sense, since we usually argue that in the ideal case, prices should represent the marginal social production costs of a certain good. Attributability of costs based on the principle of marginal costs is therefore the correct one to use for an economic evaluation. The accounting profession is moving in this direction as well with cost accounting based on the concept of a flexible cost frame work.[9]

The practical applications of this approach are directly related to improvements in the existing accounting system. Because of the decreasing marginal benefit of an improved accounting system, the theoretical "ideal" will in practice not be reached. But the current systems falls short of an adequate standard as well. It seems obvious that in a sector of such economic significance, additional resources could efficiently be spent on improving the information content of the accounting system in use. The link between economics and finance is, at the moment, still far too week.

Even a cost oriented tariff system will require some mark-up above marginal cost. This may be due to the remaining, non-attributable costs, the remaining economics of scale and a legal requirement to achieve a profit[10]. This mark-up should be a fixed proportion, rather than a variable factor for different services. In this case, tariffs and marginal cost would be proportionally related to each other.

This suggestion departs from the theory of optimal prices (or second best prices) in the Ramsey-Boiteux tradition, which diverge (inversely to the size of demand elasticity) from the marginal costs of each service. There are a number of reasons for dropping the second best pricing concept[11]. The first is that the need to estimate price elasticities introduces an additional uncertainty about the pricing structure in the system. This increases the risk of inefficient investment decision by the user of telecommunication services and the producer of associated equipment. To utilize the available technology potential fully, the tariff structure should be based on predictable principles.

The second argument relates to the fact that given the difficulty of correct demand elasticity estimation, the chances of politically motivated tariff changes are higher when tariffs are based on marginal costs and demand elasticities. This may lead to excessive cross subsidization and related efficiency losses.

...

2) A resale tax

If one starts with the proportionality between attributable costs and charges in its pure form, it is easy to see that with a shared use and resale policy, such resale will have to be taxed to make up the loss in mark-up to the PTT. The service retailer would, in addition to the rental charges for leased lines, have to pay a percentage on his own surplus, which is equivalent to the proportional mark-up of the PTT. The final users contribution for non-attributable network services are then equivalent to those which he would have paid as a direct customer of the PTT. The tax on telephone agencies can then be seen as analogous to the system of value added taxes. Their aim is also to tax the final user in the same way, independent of the way which the purchased product has taken through the production and distribution channels.

This tax would also eliminate the negative effects of cream-skimming, which we have described above. Only those telephone agencies which are just as efficient as the PTT (or better) will be able to survive with positive profits. Only they will therefore have an incentive to enter the market. The other telephone agencies which would survive under this condition are the value added networks, which in essence, produce something different from the PTT.

) Nationally Uniform Tariffs and Cost Based Tariffs

While a cost oriented tariff normally leads to the abandonment of uniform national tariffs, and an elimination of the resulting cross subsidization, such a policy may be politically infeasable. This does not necessarily mean however that the concept of a cost-based tariff or unrestricted user freedom is no longer desirable . But some adjustments are obviously necessary, to combine the benefits of increased user freedom with the political necessity to stop "deaveraging" from taking place. Two policies shall be considered here.
In first and most extreme case, we assume that the demand for uniform national tariffs is paramount ; even in the face of widely diverging regional costs. In this case, one has to weight the regional and local marginal costs for each service by their relative demand. The resulting, "average marginal cost" of that particular service will then be treated just as before, including the mentioned resale tax. This resale tax would then, in addition to the mark up for non-attributable cost, entail a subsidy to structurally disadvantaged regions. While this is a significant departure form the principle of a cost based tariff, it would at least lead to an equalization of marginal costs between services. In addition, it is an exception by political necessity, which should not hinder the PTT from pursuing a policy in which the price signals reflect costs as close as possible.

The second case is just a variation of the first, namely a necessity for the PTT to internally subsidize other services (for example the mail side of business) as we observe in a number of countries. This requires the correct calculation of the cross-subsidization involved for each particular service, again not an easy task. But this may well pay for itself in long term efficiency gains, rather then a continuous restriction of user freedom to its current level.

There are of course further combinations possible, of politically necessary cross-subsidization on one hand and increased user freedom on the other. But different requirements require different solutions, some of them administratively more cumbersome then others. The efficiency costs of maintaining cross-subsidization and restricting user freedom should therefore be kept in mind, so that eventually a direct subsidity program may overcome such cumbersome arrangement.

4) Cost oriented tariffs and positive external effects.

One of the arguments for increased competition in the network is the better exploitation of the emerging technology potential due to the technological revolution in the telecommunication sector. In this connection, it is important to recognize that the different telecommunication services are not only subtitutes for each other, but that they may also complement each other. This is especially evident with respect to some of the new services currently being tested for future services, for example view data, or teletex. In themselves and in their initial trial base, they may not be very attractive for the large segment of household

...

customers, because of the limited amount of information currently stored and the small number of subscribers involved. But a combination of the view data and telex system with alphanumeric input devices and printing facilities could change this considerably. Then the private household could make better and easier use of the information provided and use it more universally with respect to its own activities. This will, on the other hand, also make the telex system (or its variants) more attractive to the business user in his dealings with the private household. This will be especially true in the service sector. Similar tendencies are emerging with respect to electronic mail.

While this example may be a bit speculative, it points towards one important aspect of a cost-oriented tariff, externalities. This means that ideally, a cost-oriented tariff should be based on the social marginal cost, not just on those of the PTT. The idea of a cost-oriented tariff is to provide the user with price signals, which reflects the total (societal) resource use of network utilization. An external effect would lead to a divergence between private and social cost. Positive external effects in the telecommunications network develope, for example through an enlargement of the network. New subscribers create therefore an external effect for other network users. This effect is especially important when the network is still very small, as it is the case with some of the newly emerging services. If one takes these external effects into account, then one should set the tariffs in emerging networks or services below cost.

This may sound like a recommendation to cross-subsidize new services from established services. But this is a rather short sighted interpenetration. A more comprehensive analysis suggests that this initial deficit should only be seen as an investment, just as it takes place with respect to new products in the rest of the economy. Initially loss making tariff should attract further demand with the effect, that future demand will be higher than without the initial loss. The size of this initial investment must of course be in relation to the future demand.

...

5) Examples of Cost Based Tariffs

It may be useful to mention a few examples of cost-oriented tariffs and to indicate some of the likely changes of a move to that direction. In this connection it may be useful to look at three areas : domestic tariffs, international tariffs and the relationship between services.

With respect to domestic (and to perhaps even a larger extend international) tariffs, the emphasis will move from reliance on the distance component to the time component. Calls will became less dependent on distance. This also reflects a move from the traditional value of service pricing approach (which is closely linked to the Ramsey-Boiteux Pricing principle), to the cost-oriented approach. Such a move is already apparent with existing technology, but will become more pronounced with the increased use of satelite transmission.

If such a move to cost based tariffs is to be accompanied by the potential of network competition (even if only of the service type), which is highly desirable with respect to increased user freedom, tariffs will have to become harmonized, so that excessive or inefficient entry is avoided. (By that, I mean that telephone agencies live on arbitrage alone, are not necesseraly more efficient and to not provide value added services). The obvious base, on which the harmonization of tariffs is to take place, especially to avoid this kind of inefficient entry is of course a cost based tariff . A second best tariff would, given our earlier arguments, not be compatable with increased user freedom and, especially the move towards integrated services. Too many restrictions would again have to be based on the user to achieve some small short run welfare gains, against the long term benefits of an improved performance of the market as a search process.

Cost-oriented tariffs, combined with service competition, should also lead to a harmonization of tariffs between services. The tariffs for each service should be in the same relation to each other as the costs of each particular service. This move too would lead to a further increase of user liberalization, which is currently restricted because of large tariff differencials. A specific example of this is, for example, the differences between dialed and leased lines, or between data and voice service, of data and telex services, which we currently observe in Europe.[12]

...

Changes would also take place with respect to the structure of the telephone tariffs. If the telephone system were operated under a cost based tariff and network competition , it is arguable that the installation fee would tend to equal the initial cost of installing a subscriber in the system (including the costs of installing the telephone instrument in his house or place of business, laying a subscriber line to his local exchange, setting up his account, etc.) ; the subscription rate would tend to equal the "customer cost" of maintaining him in the system (including the opportunity cost of capital equipment tied up, billing expenses, etc.) ; and the call charge would tend to equal the cost of making each call (including the costs of switching, transmissing, metering and billing each call[13]. This implies that there should also be a charge for the up of calls, as this implies a utilization of the network facilities as well (currently already practiced with respect to international calls, for example in the U.K. and Austria). To charge for incomplete calls would of course also be appropriate, but would be much more difficult to enforce. In this case, the PTT's may have to depart from the principle of cost based tariffs.

This last example illustrates quite clearly however, that a departure from cost based tariffs leads right away to a restriction in user freedom. PTT's specify, mainly to avoid such regenerative traffic, the ratio between the number of main lines to a PABX and the connected stations, in order to hold the number of incomplete calls to a certain level.

IV Summary and Conclusions

In this paper, we have argued for a move to a cost based tariff policy, instead of the second best pricing policy currently advocated in the literature or the value of service pricing principle currently employed by many PTT administrations. This proposal is not made in isolation, but in conjunction with advocating the possibility of network competition (either on the basis of service or facility competition) in the face of the emerging technology potential of integrated network services. While we admit that the second best pricing principle in the Ramsey-Boiteux tradition is Pareto-superior in a static world, the benefits of increased user and producer freedom in exploiting the available technology potential better do more than make up for these static welfare losses. Our proposition rests on two arguments : First, the static welfare losses tend to be small when economies of scale in the system are small, as is suggested by empirical evidence.[14]

Second, both second best pricing and value of service pricing require the introduction of user restriction. In a static world, this reduces the substitution possibilities between services for the user and at the same time raises administrative costs. In a dynamic sense, it unduly restricts the user's participation in the search process for a more efficient exploitation of the available technology potential. This not only increases the costs for the user, but also harms the PTTs (who could receive extra income from the additional traffic demand of new applications created), and society as a whole. Furthermore, the PTTs are restricted in their search for new telecommunication applications by their obligation to serve, while private users and resale agencies face a much smaller risk in this task. They can therefore experiment much easier with the available technology potential and thereby test more widely new products and market demand.[15]

In the light of these arguments, a reduction in these user restictions seems appropriate. We have tried to show that a move towards a cost based tariff is a policy which is compatible with this aim.

Such a tariff should be proportional to the attributable costs of each service. This makes sense only when a significant part of total costs is shown to be attributable. This is a highly controversial point, especially in the light of the problem facing regulatory agencies. Nevertheless, we argue that this is the case, not so much because of empirical results, but because of linking it with the concept to economies of scale. We have shown that with an "ideal" accounting system, the proportion of attributable costs to total costs is equivalent to the scale factor of the system. Empirical evidence suggests that the scale factor is not much smaller than 1, so that attributable costs will be a large proportion of total costs. It therefore makes sense to move towards a cost structure, which more or less mirrors the attributable costs for each service.

A move in this direction implies a departure from the current use of uniform national tariffs, from value of service based tariffs and from cross subsidization. If some cross subsidization is to be maintained, for example for services to regionally disadvantaged areas, then these cost based tariffs can be modified, without having to give up the benefits of increased user freedom. To cover non-attributable cost and some of these area subsidies, a proportional mark-up above the attributable costs for each service is proposed. This proportional mark-up would also form the basis of a resale policy. It would be based on the difference between the retailers tariff and his payment for leased lines to the PTT. The final user pays, in this way, the same contribution to the financing of non-attributable costs as do the direct PTT customers, (similar to the concept of value added taxes).

Examples of cost based tariffs would be a move from distance based tariffs to time based tariffs, and to peak load pricing. In the case of significant extonalities, we propose a departure from cost based tariffs in such a way that for new services where new customers create a positive externality for other network users, connection charges should be priced below costs. This is not a departure from cost based tariffs, since the initial losses associated with such a policy would be made up by future network growth and related income.

FOOTNOTES

* This paper is based on a larger study on the issue of competition in the telecommunication sector. The financial support of the German Marshall Fund is gratefully acknowledged. The initial work on this topic was carried out under a grant from the West German Monopolies Commission, in collaboration with C. C. von Weizsäcker and G. Knieps, (Die Möglichkeiten des Wettbewerbs im Fernmeldewesen, Nomos, Baden-Baden, 1981). The thoughts of and the critical discussions with the two co-authors of that report have also influenced the author in the preparation of this note.

(1) For a survey, see Littlechild, S. C., Elements of Telecommunication Economics, London and New York, 1979.

(2) See Littlechild, S. C., the Effects of Postal Responsibility and Private Ownership on the Structure of Telephone Tariffs - An International Comparison, Working Paper, 1980.

(3) See also Müller, J. The Potential for Competition and the Role of the PTTs, Telecommunications Policy, March 1981.

(4) Averch, H. Johnson, L. L., Behaviour of the Firm under Regulatory Constraint, American Economic Review, Vol. 52, No. 3, 1962, p. 1052-1069.

(5) Hayek, F., Der Wettbewerb als Entdeckungs_verfahren, Kieler Vorträge, No. 56, 1968.

(6) Posner, R., Taxation by Regulation, Bell Journal of Economics and Management Science, Vol. 2, No. 1, 1971, pp. 22-50.

(7) See Owen, B. and Braeutigam, R., The Regulation Game : Strategic Use of the Administrative Process, Ballinger, Cambridge USA, 1978, especially chap. 1.

(8) The proof is easy. Let $F(\lambda)$ be defined for given $X_1 \ldots\ldots X_n$ as $F(\lambda) = K(\lambda X_1, \lambda X_2, \ldots \lambda X_n)$. We differenciate F at the place $\lambda = 1$ and obtain $F'(\lambda) = \sum_{i=1}^{n} X_i K_i = Z$

At the same time, the relation between marginal costs ($F'(\lambda)$) and the average costs $\left\{ \frac{F(\lambda)}{\lambda} = F(\lambda) \right\}$ is given by the equation

$$r = \frac{F'(\lambda)}{F(\lambda)} = \frac{Z}{K}$$

(9) See Kilger, W. Flexible Plankostenrechnung, 7. Ed, Opladen, 1977.

(10) This is, for example the case in France, W. Germany and the U.K.

(11) As we have mentioned above.

(12) See for example OECD, Policy Implications of Data Network Developments in the OECD Area, Paris, 1980, Chap. 10, or Department of Industry, Report of the National Committee on Computer Networks, London, Oct. 1978.

(13) In all cases, of course, due allowance would need to be made for the changing pattern of costs and demands, imperfect knowledge, the costs of establishing and enforcing a differentiated tariff, etc. All these are examples of components, which must be taken into account with a move towards cost based tariffs. The currently practiced peak load pricing is of course part of this process as well.

(14) For a summary, see Charles River Inc; The Economics of Competition in the Telecommunications Industry, Boston, 1979.

(15) Even if the reduction in user restriction results in an initial reduction of PTT's income, one should keep in mind that the wider use of new telecommunications services in the commercial sector is often only a prelude to the wider applications in the household sector.

REVIEW OF
"THE VALUE OF THE FIRM UNDER REGULATION AND THE THEORY OF THE FIRM UNDER UNCERTAINTY: AN INTEGRATED APPROACH"

BENOÎT DESCHAMPS
GEORGIA STATE UNIVERSITY

As the title of his paper suggests, Professor Perrakis is incorporating within the framework of microeconomics of uncertainty several of the elements of the theory of the regulated firm. Uncertainties allowed into Prof. Perrakis' models are random demand and cost functions, and he is dealing with the Averch-Johnson[1] [1] behavior and two types of rate of return regulation, one whereby the allowed return is based on expected performance (forward looking regulation) while in the other a regulatory review is instituted when the realized return falls outside upper or lower bounds. Finally, the paper employs several approaches to valuation, some more traditional than the others.

All of this being dealt with in a single paper is impressive, and as such Professor Perrakis' contributions should be recognized.

The first task undertaken in Professor Perrakis' paper is to analyze the validity of Modigliani and Miller [28, 29, 30] leverage propositions to the firm subjected to rate of return regulation. His first result is basically a re-statement (using arbitrage valuation operators) of Jaffee and Mandelker's [19] conclusion that in general

$$V_L = \frac{\bar{N}_L(1-t)}{\rho^*} + t\ D \neq \frac{\bar{N}\ (1-t)}{\rho} + t\ D \tag{1}$$

where ρ^* is the relevant capitalization rate for the unlevered stream $\bar{N}_L$.

The paper then turns to analysing the implication of that conclusion on the celebrated Modigliani and Miller (M&M) formula for the return on a share of a levered firm:

$$\rho_L = \rho^* + (\rho^* - r)\ (1 - t)\ L \tag{2}$$

Unfortunately, that part of the analysis is somewhat circular in the sense the initial premise is basically that the RHS of M & M formula (2) is equal to itself (see appendix). Further, the condition for M & M formula to hold is obtained after substituting into it

$$\frac{\bar{N}_L}{V_L - t\ D} = \frac{\hat{N}}{V} \tag{3}$$

which violates the conclusion already reached by Jaffee and Mandelker. However, this section of the paper still adds to Jaffee and Mandelker's analysis since it allows

[1] References are numbered according to Professor Perrakis' article.

for different levels of capital and inputs, making it more unlikely that N_L and N are ~~perfectly correlated~~ strictly proportional as in (3).

As far as the treatment within the Capital Asset Pricing Model of the Averch-Johnson type of investment behavior is concerned, that section of the paper would merit some expansion. First, the CAPM is not vital to the section since it does not explicitly appear in the solution (we only have the derivative of V with respect to output price); therefore, the CAPM could be omitted. Second, the CAPM would have been most useful to analyze the exact relationship between the return on a share of a levered regulated firm with that of the return of an unlevered firm. This type of analysis involving selection of different levels of inputs, output and leverage has been done by Hite[2], and it could have provided interesting insights into the investment behavior of the regulated firm.

The last two sections of the paper dealing with backward-looking regulation, could be the subject of a separate article. They employ a different valuation framework and leave aside issues involving different levels of leverage. These sections constitute an interesting application to the regulated firm of the theory of valuation using options.

It is however worth noting that one of the problems raised in multiperiod valuation under uncertainty is that the level of capital stock may change from period to period and that the contingent claim approach employed in the paper assumes that problem non-existent.

[2] See Hite, G.L.: "Leverage, Output Effects, and the M-M Theorems", Journal of Financial Economics 4:2:177 (March 1977).

Appendix

Let:

$$\rho_0 + (\rho_0 - r)\ (1 - t)L = \rho_0 + (\rho_0 - r)\ (1 - t)\ L$$

$$= \rho_0 + \rho_0\ (1-t)\ L - r(1 - t)\ L$$

$$= \rho_0[1 + (1 - t)L] - (1 - t)\ r\ L$$

Let L = D/S

$$= \rho_0\ \frac{[S + (1 - t)\ D]}{S} - (1 - t)r\ \frac{D}{S}$$

Assume $\frac{N_L\ (1 - t)}{S + (1-t)D} = \frac{N(\ 1 - t)}{V}$:

the firm's expected operating income differs only by a scale factor from the unlevered firm's and this scale factor is [S + (1 - t) D] /V (this is M & M conclusion). Then:

$$\rho_0 + (\rho_0 - r)\ (1 - t)\ L = \frac{\frac{N_L\ (\ 1 - t)}{[S + (1-t)D]}}{\frac{N\ (1 - t)}{V}} \cdot \rho_0\ \frac{[S + (1 - t)\ D]}{S} - (1 - t)\ r\ \frac{D}{S}$$

$$= \frac{N_L\ (1 - t)}{N\ (1 - t)} \cdot \rho_0\ \frac{V}{S} - (1 - t)\ r\ \frac{D}{S}$$

$$\rho_0 + \rho_0\ (1 - t)\ \frac{D}{S} - r(1 - t)\ \frac{D}{S} = \frac{N_L}{N}\ \rho_0\ \frac{V}{S} - (1 - t)\ r\ \frac{D}{S}$$

$$\frac{N_L}{N}\ \frac{V}{S} = 1 + (1 - t)\ \frac{D}{S}$$

REVIEW OF
"FINANCING AND INVESTMENT BEHAVIOR OF THE REGULATED FIRM UNDER UNCERTAINTY"

BENOÎT DESCHAMPS
GEORGIA STATE UNIVERSITY

Professors Berkowitz and Cosgrove develop a model of the investment and financing decisions of a rate of return regulated firm where bankruptcy, and its associated costs, are permitted. Within the framework of the paper, one would have expected the model to rest on the solution of the following maximization model:

MAX Firm current value

Subject to: The regulatory constraint, as stated by regulators
the chance constraint on bankruptcy risk tolerance

Although the basic formulation of the model is along these lines, the regulatory constraint is ingeniously stated so that the difference between allowed (regulated) earnings[1] and expected (regulated) earnings is set equal (when the constraint is binding) to the level of physical capital times the difference between the allowed return and the average cost of capital (eq. 7).

Given the fact that the firm value should be nothing (in the current framework) but the ratio of expected earnings divided by the cost of capital, a question is raised regarding the possibility of formulating that constraint: the solution value of the firm should be known in order to estimate the cost of capital employed in the constraint.

A related issue is the fact that the paper is not explicit as to which variables are treated as constant in the value maximization problem. A useful addition would be an appendix showing how the first-order conditions (13 a to d) were obtained. For instance, although the cost of debt as perceived by the regulators is allowed to vary with leverage, it is by no means clear if the actual cost of debt and equity would vary with leverage. Clearly, if they are not, there is inconsistency in the paper.

A third point is the confusion between "money" capital and "physical" capital.[2] I share Meyers' approach that the price of physical captial needs to be distinguished from the amount of funds required to pay for it. Because the authors did not distinguish between money and physical capital, they introduced a dichotomy between their definition of income and their valuation formula, whereby what is discounted should be income (recovery of capital can be part of income, but then it has to be included in the definition of income).

[1] "Regulated" earnings is here defined as net profit after taxes plus interest on debt.

[2] For a discussion at length of this point, see D. Vickers: The Theory of the Firm: Production, Capital and Finance. McGraw-Hill Book Company, New York 1968.

Similarly, the cost of capital expression (eq. 5) is mis-stated[3]: it does not account for taxes and it is based on "book" values of debt and equity. It is also worth noting that while $\bar{\delta}$ is the selected level of leverage, δ (without a bar) is the level of leverage considered optimal by regulators. As a result, $\bar{\delta}$ and not δ should be a decision variable in the maximization problem.

Even if the model was properly specified, some of the conclusions reached in the paper would need to be either corrected or substantiated. For instance, even if the selected debt-equity mix is lower than optimal, it does not make less problematic the controversy between Gordon (1967) and Elton and Gruber (1971) over whether or not the tax benefits of debt are regulated away. Similarly, if the marginal benefit of an extra dollar of debt is zero, it does not mean that a corner solution of 100% debt has been reached, but that additional debt over and above the optimum is not worthwhile. Finally, when considering regulation based upon embedded costs, one cannot necessarily assume that the embedded cost of debt is below marginal cost.

[3] For a discussion of different ways to measure cost of captial, see T.J. Nantell and C.R. Carlson "The Cost of Capital as a Weighted Average," Journal of Finance 30:5:1343 (December 1975).

DISCUSSION OF
"TAXES, FINANCING AND INVESTMENT FOR A REGULATED FIRM"

BENOÎT DESCHAMPS*

Professor Bernstein's paper is one of the few[1] dealing with the application of control theory to the problem of selection by firms of alternative financing means. In that sense, it is exploring new research paths since most studies have until now employed to so-called static or, at best, comparative static framework to analyse financing and investment choices of business firms.

A second challenge in Professors Bernstein's paper is that it deals with the financing behavior of the regulated firm, an area that has received much attention in the last 15 years, but which usually brings additional complexities from the celebrated Averch - Johnson [1] investment behavior of regulated firms, as well as the intricating relationship between the regulatory constraint, and the effect of debt financing on the value of the firm.

This paper presents a theoretical model and, as such, it does not need to be an exact representation of reality. Depending upon the stage in the research process at which the theorectical model stands, some of its assumptions may have little relation with reality. One such assumption in Professor Bernstein's paper is that investors either have perfect foresight of the future (perfect certainty) or they are risk neutral. Even is we do not criticize that assumption, other assumptions of the paper ought to be consistent with it. Another underlying assumption is that capital markets are nearly perfect[2] in the sense usually

*Associate Professor of Finance, Georgia State University

[1] See Krouse [4], Mehta [5], Elton and Gruber [2] among others.

[2] See Fama and Miller [3]. Allowed financial markets imperfections include a systematic effect of taxes so that securities are traded on an after taxes basis (see also [3, p. 174], and institutional constraints on the role of firms in the financial markets.

employed by financial theory. Together, these two assumptions imply that all securities are traded on the basis of a same and identical expected return.

Usually, in that context, personal and corporate taxes are ignored so that is is a matter of indifference to choose between debt or equity financing. In this paper, debt financing is made a pertinent variable by introducing homogeneous and constant tax rates for investors so that the securities are traded on an after-tax basis. Together with corporate taxes, the effect for the investor is:

- Single taxation of interest income (or return)
- Double taxation of dividend income (or return)
- More than single but less that double taxation of capital gain income (or return)

If the maximand in Professor Bernstein's paper had been maximization of (current) total firm value (i.e. debt plus equity) instead of maximization of (current) stock value, a properly specified model allowing shareholders to buy new bond issues would have probably yielded at situation where the firm would provide as much return it can in the form of interest, then in the form of capital gians and, at last, in the form of dividends. In Tables 1 and 2, new bond issues do not always enter (non-zero) in the solution because the maximand does not really allow shareholders to buy bonds. Furthers, if no dividends are paid by a non-regulated firm, it is likely to be simply because capital gains provide more after-tax dollars to the investor.

However, the fact that no dividends would be paid by the firm in the final solution introduces some kind of irrational behavior from the part of investors. Unless Professor Bernstein is developing an entirely new theory of stock valuation, he has to follow the lines first suggested by J. B. Williams [7][3]: the value of one share of stock is the discounted value of all future

[3] and extended by Miller and Modigliani [6]

per share cash dividends, while the aggregate value of currently outstanding shares is the discounted value of all future aggregate cash dividends minus the proceeds from future stock issues.

The maximand (eq. 9) actually has that form (on an after-tax basis), but since $s \geq 0$, all but two of the solutions shown on Table 1 imply that dividends are zero and that therefore the value of an unregulated firm's stock is zero! At that juncture, one is left to wonder whether the model was properly specified to begin with. Actually, these zero solutions were obtained through the assumption that after-tax capital gains are equivalent to after-tax cash dividends. However, that assumption can only be maintained if capital gains originate[4] from an increase in anticipated dividends, which are in the solution zero, regardless of time.

The second point of this discussion regards the regulatory framework. The manner a company is regulated in a theoretical model does not need to be actually observed: the model may seek to develop new approaches for regulation. The framework adopted in this paper is rather unusual: it is neither earnings or selling prices that are regulated but rather the ratio of "retained earnings" to total assets. However with a constant debt r (which is arrived at in the steady-state solution), this turns out to be identical to "rate of return" regulation whereby dividends are treated as an expense instead of a remuneration of shareholder's money capital.

In all logic, within that framework, if dividends are treated like an expense by regulators, firms should seek to provide as much dividends they could to their shareholders, so that eq. 5 should be binding and $E = 0$ (all new investments are externally financed). In the paper, this is more like the case of the unregulated firm since effective regulation would force

[4] Otherwise, the value of a stock may be determined just like in a beauty contest, the most beautiful women is selected according to the tastes of <u>that</u> contest's judges.

retention to be exactly equal to the "allowed retention ratio". Actually, in the framework developed in the paper, we should expect regulation to be never "effective" in the sense that firms would pay more dividends to make it non-effective unless the tax rates are selected so to make regulation worthwhile.

Finally, to discuss some of the conclusions in the paper, one has to note that when the allowed retention rate is set below the economic depreciation rate, the value of currently outstanding shares should in the long run tend towards zero: either the firm would be self-liquidating or the proportion of (after-~~time~~ tax) before interest earnings taken up by new bondholders or shareholders will eventually reach 100%.

References

[1] Averch, H. and Johnson, L.,"Behavior of the Regulated Firm", American Economic Review (December 1962), Vol. 52, pp. 1052-1069.

[2] Elton, E. J., Gruber, M. and Lieber, Z., "Valuation, Optimum Investment and Financing for the Firm Subject to Regulation", Journal of Finance, Vol. 30, no. 2 (May 1975), pp. 401-425.

[3] Fama, E. and Miller, M. H., The Theory of Finance, New York: Holt, Rinehart and Winston (1972).

[4] Krouse, C. G., "Optimal Financing and Capital Structure Programs for the Firm", Journal of Finance, (December 1972), Vol. 27, no. 5, pp. 1057-1071.

[5] Mehta, D. R., "Impact of Financial Policies on the Valuation of the Firm: A Control Theoretic Approach", paper presented at the XXII International Meeting (1975), The Institute of Management Sciences, Kyoto, Japan, in the session on Applied Optimal Control.

[6] Miller, M. H. and Modigliani, F., "Dividend Policy, Growth and the Valuation of Shares", Journal of Business, Vol. 34, no. 4, (October 1961), pp. 411-433.

[7] Williams, J. B., The Theory of Investment Value, Cambridge, Massachusetts: Harvard University Press, 1938.

Reply to Professor Deschamps Comments on:

"FINANCING AND INVESTMENT BEHAVIOUR OF THE REGULATED FIRM UNDER UNCERTAINTY"

M.K. Berkowitz and E.G. Cosgrove

The comments made by Benoit Deschamps indicate a need on our part to clarify some of the subtle and potentially confusing issues in our paper. We welcome the opportunity to do just that.

In his first comment, Deschamps argued that the solution value of the firm should be known in order to estimate the cost of capital employed in the constraint. Indeed, the valuation equation might be written alternatively as the discounted expected earnings where the rate of discount is the cost of capital. That is,

$$\text{(A-1)} \qquad V = \frac{E(\tilde{Y})-\lambda \mathrm{cov}(\tilde{Y},\tilde{Y}_M)}{1+\rho} = \frac{E(\tilde{Y})}{1+r}$$

$$\text{where} \quad r = \rho+\beta(\bar{R}_M-\rho)$$

It is generally recognized that the above valuation expressions are equivalent. Moreover, r is the same opportunity cost of funds that appears in our regulatory constraint. The problem, however, is not solved in a stepwise manner, as suggested by the discussant. To the contrary, the decisions taken by the managers of the firm yield a particular risk structure for the firm and hence opportunity cost of funds - r. That is, the determination of r and the decisions taken by the firm are simultaneously determined in both the valuation equation and the accompanying constraint. The assumption is made, furthermore, that regulation is continuously enforced so that (s-r) remains constant irrespective of changes in the firm's decisions which alter the risk of the firm. This is consistent with prior work in the area, e.g. Averch-Johnson, etc.

As for the second point, we have not included the derivation of the first-order conditions in the interest of parsimony and because the derivation was a straightforward exercise. We continue to believe that no useful purpose would be served by simply adding an appendix so as to expand the paper by some 5-8 typed pages. There is no inconsistency in the paper since the actual costs of debt and equity as well as the allowed costs vary with leverage.

The discussant does not seem to appreciate that the rental price of capital in a competitive market includes the purchase price, depreciation, and financing costs (return on investment) associated with the purchase and use of the physical capital stock. To simply subtract the interest charges (iB) from the economic profit of the firm would be deducting the non-ownership opportunity cost of funds twice. Indeed, Meyer (1976) followed Vickers (1968) and both were incorrect. The discussant would do well to refer to Takayama (Mathematical Economics, 1974) and associated references therein in order to understand the distinction in our paper between physical and financial capital.

The fourth point raised by Deschamps is that the cost of capital expression in our paper is mis-stated. Clearly, this is not the case. To demonstrate, let us assume for simplicitly the absence of bankruptcy. It follows that the expected returns to shareholders and bondholders after payment of taxes $E(\bar{Y})$, is:

$$\text{(A-2)} \qquad E(\bar{Y}) = (1-\tau)\,[E(\bar{\Pi})-\bar{i}\bar{B}] + \bar{i}\bar{B} \ ,$$

which can be written alternatively as

$$\text{(A-3)} \qquad E(\bar{Y}) = \bar{k}\bar{S} + \bar{i}\bar{B}$$

where $\bar{k}$ is the after-tax return to shareholders and $\bar{i}$ is the before-tax return to bondholders. Rearranging (A-2),

$$\text{(A-4)} \qquad E(\bar{Y}) = E(\bar{\Pi})(1-\tau) + \bar{i}\tau\bar{B}$$

It follows from (A-3) and (A-4),

$$E(\bar{\Pi})(1-\tau) = \bar{k}\bar{S} + \bar{i}\bar{B}(1-\tau) \ , \quad \text{or}$$

$$E(\bar{\Pi}) = \left(\frac{\bar{k}}{1-\tau}\right)\bar{S} + \bar{i}\bar{B} \ , \quad \text{or}$$

$$\text{(A-5)} \qquad E(\bar{\Pi}) = \bar{\alpha}\bar{S} + \bar{i}\bar{B} \qquad \text{where } \bar{\alpha} = \frac{\bar{k}}{1-\tau} \ .$$

The expression in (A-5) clearly reflects the before-tax expected profits on both sides and is therefore consistent. A similar derivation for allowed profit will yield the regulatory constraint adopted in our model.

Furthermore, our equations (5) and (7), in the text, which define the cost of capital and related cash flows are identical to equations (2) and (2a) in the Nantell and Carlson (1975) article referred to by Deschamps. The equivalence of the various forms of the cost of capital are rationalized in the Appendix to the N-C paper and therefore require no change in our presentation.

Deschamps is totally confused in his description of $\overline{\delta}$ and δ (without the bar). We define δ as the selected level of leverage on p.3, and clearly state that $\overline{\delta}$ is the regulator's perceived optimal debt equity level (p.6). The distinction can be further noticed in going from equation (7) to (8). It is quite evident from the above that δ (without the bar), not $\overline{\delta}$ as Deschamps proposes, is the appropriate decision variable in the maximization problem, and that is what we did.

As for one of Deschamps final remarks, the following discussion expands our comments on the Gordon (1969) Elton-Gruber (1971) controversey. Gordon claimed that regulators adjust prices in such a way that the after-tax earnings plus interest, for a regulated firm, is a constant, independent of the debt equity ratio — hence he claimed that $dV/dB = 0$. Elton and Gruber were able to demonstrate that Gordon's model was misspecified, and when Gordon's definition of regulation was employed

$$\frac{dV}{dB} = \tau\frac{(\rho-i)}{\rho} \neq 0 ,$$

where ρ = cost of capital for an unlevered firm

i = interest rate on the firm's debt

τ = tax rate

While additional debt had a <u>lower</u> impact on firm value than in the unregulated case (where $dV/dB = \tau$) , the aggregate effect was <u>not</u> totally eliminated as Gordon claimed.

We found that the firm has an incentive to select an optimal debt-equity <u>below</u> the regulator's optimal level, regardless of the regulator's effectiveness in regulating away the tax benefits from excess debt. In as much as the value maximizing firm does not find "that additional debt over the optimum to be worthwhile" (to quote Deschamps), then dV/dB must be quite small (zero?) and in that sense the controversy between Gordon and Elton-Gruber is much less problematic.

We wish to thank Deschamps for his insights on our conclusion regarding the marginal benefits of debt in the absence of regulation and bankruptcy (which are clearly the MM corporate tax case). His conclusions suggests that we may have been the first to demonstrate the existence of an optimal capital structure in the above case without having to invoke either agency costs, bankruptcy costs, personal taxes, signalling theory, corporate tax credits, etc. There is, however, no underlying justification (e.g. market imperfection) for assuming anything but a corner solution in our model.

Finally, although Deschamps may be technically (and trivially) accurate in arguing against our claim that the embedded cost of debt is below the marginal cost, our conclusion rests on casual observation of interest rates over say the last 15 years!

DEMAND ESTIMATION

PROBLEMS AND ISSUES IN MODELING TELECOMMUNICATIONS DEMAND

Lester D. Taylor
University of Arizona *

I. INTRODUCTION

It might be thought that 100 years would be sufficient for the revolution in communications launched by the words, "Mr. Watson, come here!" to have run its course, but this is not the case. For now five years into its second century, the telephone industry appears more revolutionary than ever, and most, if not all, of the turmoil can be traced to competition that has been:

-- facilitated and fueled by rampant technological change,

-- spurred by sharply higher energy and transport costs and the emergence of efficient, greatly expanded telecommunications networks outside of North America, and

-- actively encouraged in the U.S. by regulators and the courts.

And, in the midst of all this, the U.S. Congress is still trying to revise the Communications Act of 1934 in a way that will establish the ground rules for a greatly restructured telecommunications industry.

One cannot analyze the directions that the telecommunications industry is likely to take in the years ahead without an understanding of the structure and determinants of telecommunications demand. Let me give an example. Competition is now a fact of life in the toll, private-line, and terminal-equipment markets, and "contributions" from these markets, which have for many years been used to subsidize basic service to residential customers, would appear to be imperiled. As a consequence, pressures are mounting to raise basic-service rates and, in addition, to begin charging for local calls on a measured basis. However, a recent study by Charles River Associates [Meyer *et al.* (1979)] of competition in the telecommunications industry has concluded that competition in the toll market will not necessarily exert upward pressure on basic-service rates because of an elastic demand for toll calls: decreases in toll rates triggered by competition will lead to an actual increase in toll revenues. This is a questionable conclusion in my view, but it clearly points up to the importance of knowledge of telecommunications price elasticities of demand.

The focus in this paper is accordingly, on telecommunications demand. My intent is to provide a brief, but nevertheless fairly comprehensive, review of the present state-of-the-art in telecommunications demand modeling. In doing this, I shall draw heavily on my recent monograph on the subject [Taylor (1980)]. I shall begin the discussion in Section II with a brief review of the basic characteristics of telephone demand. This will be followed in Section III by a summary and critique of the existing econometric literature on telecommunications demand, and, then, in Section IV by a discussion of the problems which in my opinion are most in need of research.

II. THE THEORY OF TELEPHONE DEMAND [1]

The characteristics of telephone demand that most set it apart from the demand for other goods and services include:

(a) A distinction between the demand for access to the telephone network and the demand for the use of the network once access has been acquired.

(b) The dependence of the demand for access on the demand for use.

(c) The presence of access (or network) and call externalities which impart public-good aspects to the telephone network.

(d) The importance of option demand in determining the demand for access.

Determination of the Demand for Access

The point of departure in modeling telephone demand is the distinction between the demand for access and the demand for use, which follows from the fact that one must be connected to the telephone network before the network can be used. The purchase of access can accordingly be viewed as the purchase of the right to make and receive calls. Thus, use is conditioned on access, yet access is in turn dependent on the benefits that arise from use: for if the net benefits from use are not at least as great as the purchase price of access, access will not be purchased.

[1] The discussion in this section follows closely the presentation in Taylor (1980). Those interested in an analytical treatment of the topic are referred to Chapter 2 of my monograph.

For most households, the net benefits from use exceed the price of access, probably by a comfortable margin, and consequently, we observe that about 95% of U.S. households have a telephone.

The dependence of the demand for access on the demand for use can be easily illustrated graphically, hence let us turn our attention to Figure 1b, which shows the demand for telephone calls as a function of the price of a call. At a price of π_o, the number of calls that would be made would be q^o at the cost of $\pi_o q^o$, represented by the rectangle $0\pi_o dq^o$. The net benefits associated with these q^o calls are given by the triangle $\pi_o cd$, which is the consumer's surplus associated with the q^o calls. Denote this quantity by S_1. It is clear that access to the telephone network will be demanded if S_1 is greater than r_o, but not if S_1 is less than r_o.

In panel a of Figure 1, the price of access is represented on the vertical axis, while the consumer's surplus S_1 is represented by the spike at α along the horizontal axis. In this case, S_1 is assumed to be greater than r_o, so that access to the network is in fact demanded. However, assume now that there is a second consumer whose consumer's surplus from use of the network, S_2, is represented by the spike at β in Figure 1a. For this consumer, the net benefits from use are less than the price of access ($S_2 < r_o$), so that access will not be demanded.

More generally, let us assume that we have a population of M potential subscribers to the network. How many of these potential subscribers will be actual subscribers? To answer this question, let us consider Figure 2b, which refers to the aggregate demand for telephone calls for all M

Figure 1a

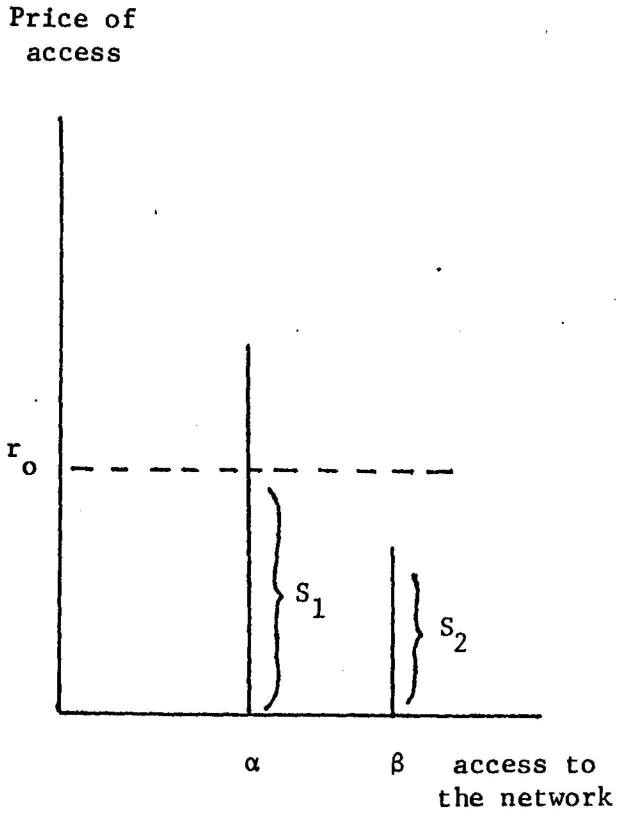
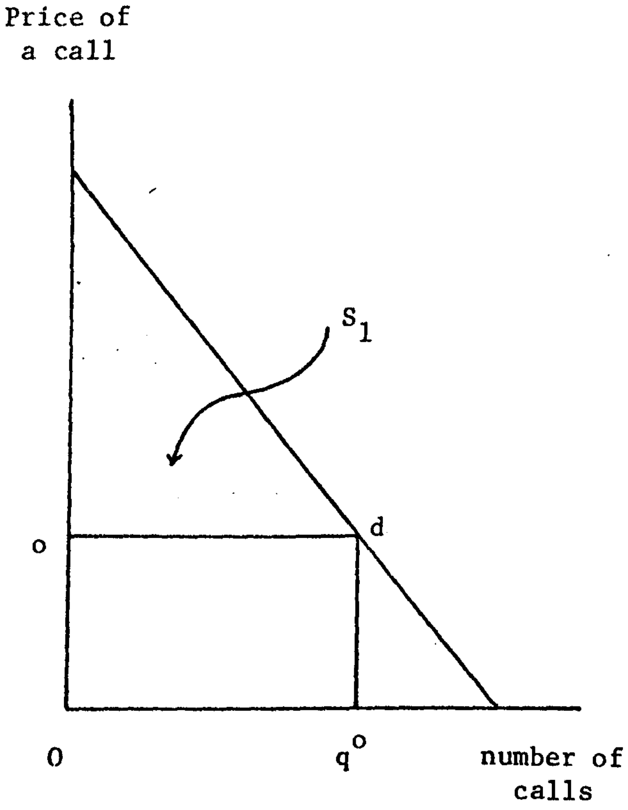


Figure 1b



potential subscribers. This demand function is derived on the assumption that everyone belongs to the network, and is obtained as the horizontal summation of the M individual demand functions. At a price of π_o per call, we see that Q_o calls will be made, where

$$(1) \qquad Q_o = \sum_{i=1}^{M} q_i^o ,$$

q_i^o being the number of calls made by consumer i.

Let the net benefits (i.e., the consumer's surplus) associated with q_i^o for the i^{th} consumer be denoted by S_i, and assume that the S_i are ordered in ascending size, so that $S_1 > S_2 > \ldots . > S_M$. These net benefits represent the willingnesses-to-pay for access by the M consumers in the population, and are described by the step function in Figure 2a. The number of consumers is measured along the horizontal axis in this figure, while the net benefits from use are measured on the vertical axis. As before, assume that the access purchase price is r_o. At this price (measured on the vertical axis), we see that to the left of the point N_o on the horizontal axis, the net benefits from belonging to the network are greater than r_o, whereas to the right, they are less than r_o. At N_o, we have $S_i = r_o$, so that consumer N_o is the marginal subscriber, and the telephone system consists of N_o subscribers. The Q_o calls in Figure 2b will consequently be made by these N_o subscribers. On the other hand, if the price of access were $r_1 > r_o$, the number of subscribers would be reduced to N_1. The Q_o calls would now be made by these N_1

Figure 2a

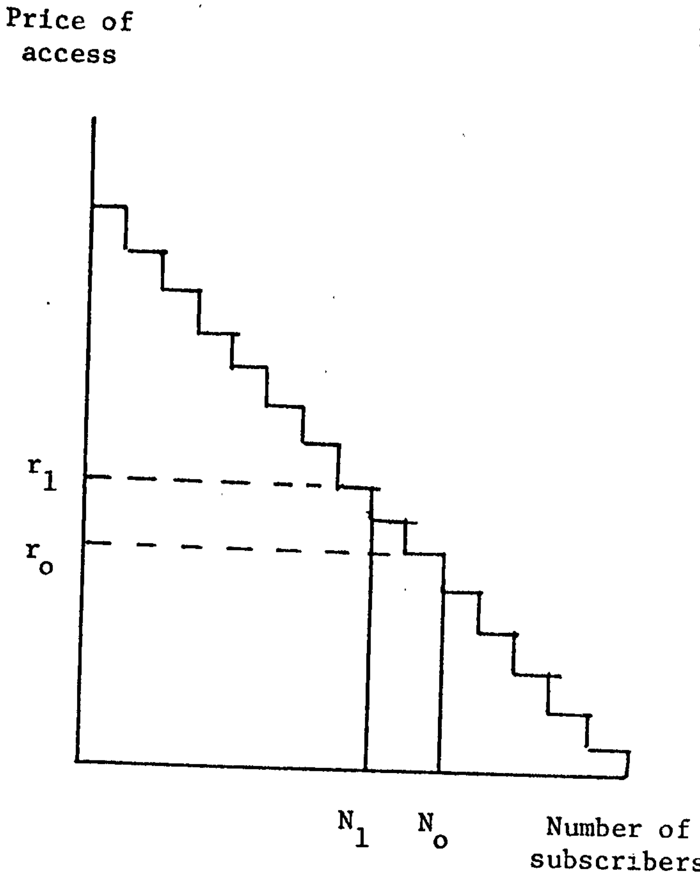
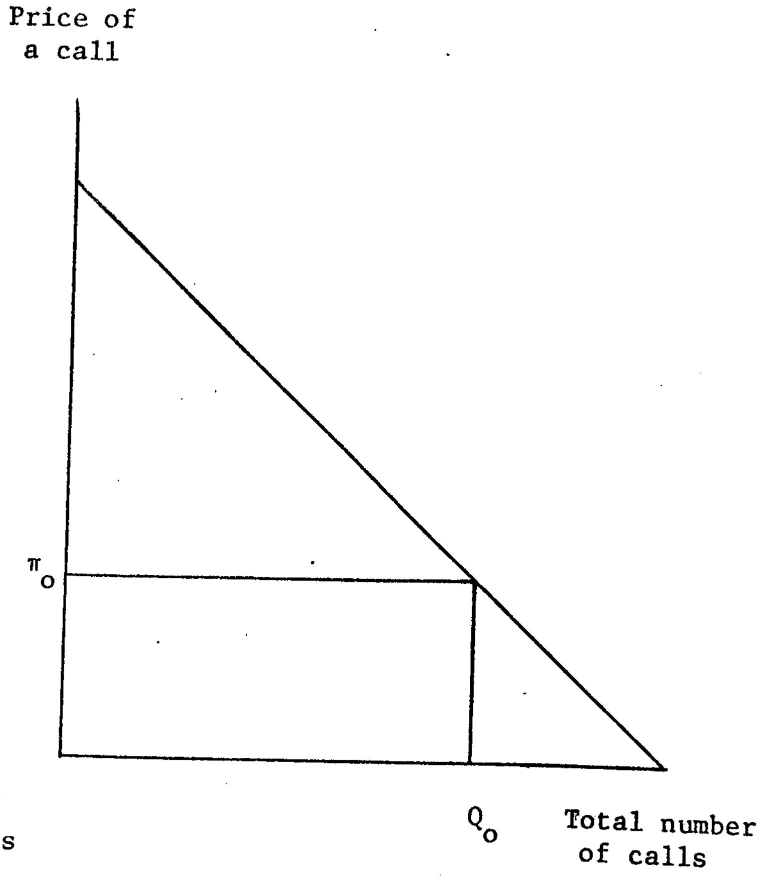


Figure 2b



subscribers, and consumer N_1 would be the marginal subscriber. The previously marginal subscriber N_o would no longer belong to the network.

It must be emphasized that the total demand for access depends on the price for calls as well as the price of access. An increased charge for calls would reduce the net benefits from use for all consumers, and since this would decrease the willingnesses-to-pay for access, the aggregate demand for access would shift to the left. This is illustrated in Figure 3. At a price of π_o per call, Q_o calls are demanded, and the aggregate demand for access is given by the step function labeled D_o in Figure 3a. If the price per call is increased to π_1, the number of calls will be reduced to Q_1, and the aggregate consumers' surplus will be reduced by an amount equal to the hatched area. This decrease in willingnesses-to-pay shifts the aggregate demand for access to the left, as indicated by the step function labeled D_1 in Figure 3a. With an unchanged access price of r_o, the number of consumers demanding access to the network is seen to be reduced to N_1 from N_o.

The Network and Call Externalities

In developing the dependence of the demand for access on the demand for use, we have ignored the complications caused by the network and call externalities, option demand, and the opportunity cost of time. The network and call externalities will be dealt with next. The network (or access) externality arises from the fact that when a new subscriber joins the network, there is now one more telephone that can be reached. This makes the network more valuable to existing subscribers, and increases

Figure 3a

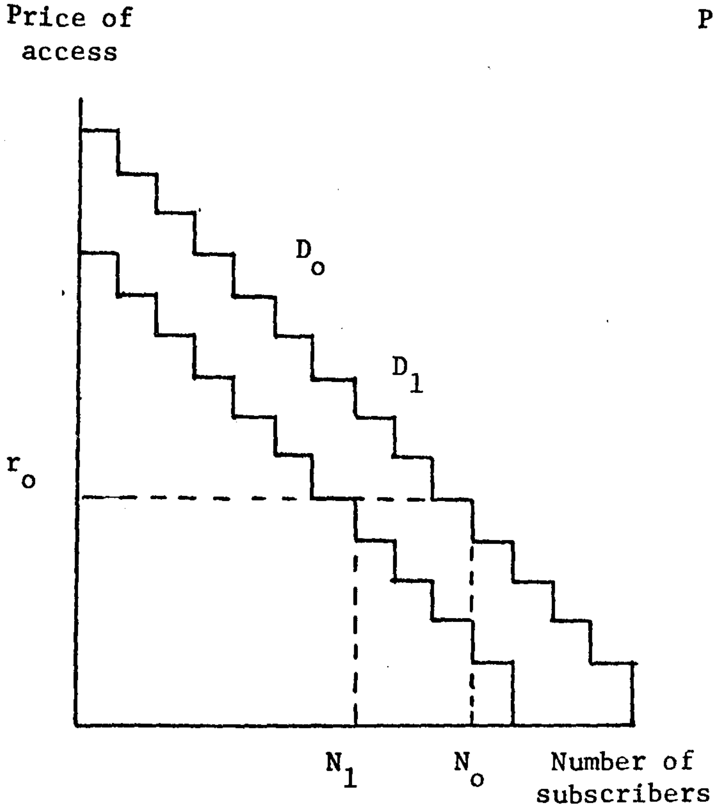
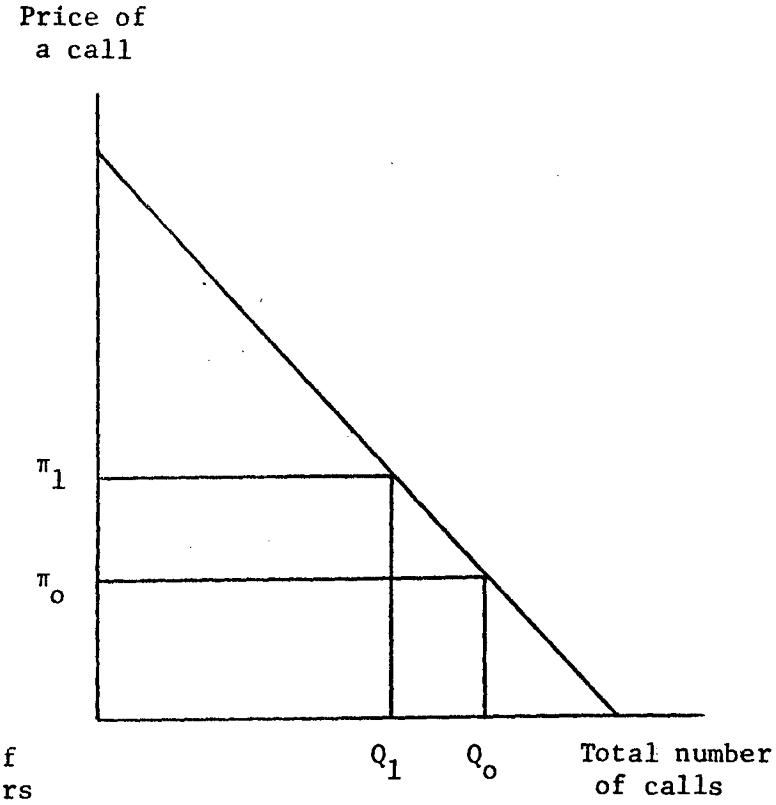


Figure 3b



their willingnesses-to-pay to remain in the system. As a consequence, consumers will be willing to pay more to join a large system than to join a small system. From this it follows that the aggregate demand function for access in a large system will lie to the right of its location in a small system.

The access externality has two important implications. The first of these, which has been analyzed extensively by Artle and Averous (1973) and Rohlfs (1974), refers to the equilibrium size of the telephone network. Because of the externality, the equilibrium size of the system will be larger than what it would be in the absence of the externality. Less clear, however, is the fact that the equilibrium size of the system may not be unique. One particularly interesting possibility is where there are two equilibria, one with a small number of telephones and the other with a much larger number of telephones. Suppose that the equilibrium is initially at the lower value, and let there be a displacement from this equilibrium, triggered (say) by an increase in income of sufficient size to cause a previous nonsubscriber to become a subscriber. In this situation, the system may not remain at the lower equilibrium, but, because of the network externality, could increase in size until it reaches the higher equilibrium. Such growth would be endogenous because it could occur in the absence of any further changes in income.

The mechanism is simple to describe: As the number of telephones increases, the network externality makes belonging to the system more valuable, which causes nonsubscribers previously on the margin to become subscribers. The size of the system would accordingly increase, thereby becoming more valuable to belong to, and once again leading a new group of nonsubscribers to join the system. Such endogenous growth would continue to the point where the last round of new subscribers failed to make belonging to the system sufficiently valuable to cause a marginal group of nonsubscribers to subscribe. Endogenous growth of this type is probably a number of years in the past in North America, but may be currently in progress in countries such as France and Saudi Arabia.

This second important implication of the access externality involves a normative question and relates to the way that access to the telephone network should be priced. The problem is as follows. An individual consumer makes the decision to join the telephone network strictly on the basis of the private benefits to him. However, in joining the network, a new subscriber confers a benefit on all existing subscribers, so that the total social benefit of the decision to join are greater than the private benefits. Consequently, it follows that if access were priced according to marginal cost, the equilibrium size of the system would be smaller than what would be socially optimal. This is because the price charged would be equal to the marginal private benefit of belonging to the system, but less than the marginal social benefit. Social optimality consequently requires an access charge that is below marginal cost. How far below obviously depends upon the quantitative importance of the externality, and is therefore an empirical question.

In contrast to the access externality, the call externality arises from the fact that a completed call requires the participation of a second party, and refers to the benefit that is conferred on this party by the person making the call. This benefit is treated as an externality because, except for collect and inward WATS calls, the cost of a call is borne by the caller. While there are undoubtedly many instances where a called party does not feel benefited by a call, the externality is clearly positive on balance, and adds to the willingness-to-pay to belong to the telephone network. Because of this, the equilibrium size of the network will be larger than what would be the case in the absence of the externality.[2]

Option Demand

A further complication in analyzing telephone demand arises from the fact that benefits are associated not only with completed calls, but also with calls that may not be made. When an individual subscribes to the telephone system, he in effect is purchasing options to make and receive calls. Some of these options will be exercised with certainty, while others will be made only randomly. This is because many calls are only made contingent upon particular states of nature whose realizations are random, and therefore not known at the time that access to the telephone system is purchased. Calls of an emergency nature, as for fire, police, or ambulance, are obvious cases in point.[3] However, compelling urgency is not the only determinant, for options may be purchased because preferences themselves are random. We shall have more to say about this in a

[2] Note that, unlike the access externality, the call externality does not lead to a pricing problem because the benefits conferred can be uniquely attributed, and can therefore be captured in the access charge.

[3] The "hot line" between the White House and the Kremlin provides another example.

moment.

Option demand, which was first discussed by Weisbrod in a provocative article published in *The Quarterly Journal of Economics* in 1964, has figured prominently in the economics of irreplaceable natural resources and areas of natural beauty. As described by Krutilla (1967):

> [Option] demand is characterized as a willingness to pay for retaining an option to use an area or facility that would be difficult or impossible to replace and for which no close substitute is available. Moreover, such a demand may exist even though there is no current intention to use the area or facility in question and the option may never be exercised. [Krutilla (1967), p. 780].

While the telephone network may not be viewed by everyone as a thing of great beauty, it is clear that option demand in the sense just described is an important component of telephone demand. To fix ideas, let us assume that during a given period of time (say a month), a consumer is willing to pay something for the option to make R calls. These are in addition to the calls that the consumer knows with certainty will be made. Let θ denote the proportion of these calls that will in fact be made. Assume that both R and θ are known numbers for the consumer. Assume, further, that the expected value of the options that will be exercised, θR, is included in q^o (as defined earlier), so that the net benefits from these calls are already included in the consumer's surplus associated with q^o.

On the other hand, the net benefits from the $(1 - \theta)R$ options that are not being exercised will not be represented in the consumer's surplus associated with q^o, thus this measure of the net benefits from using the telephone system understates the amount that the consumer is actually willing to pay in order to have access to the system.[4] Let ω denote the benefit yielded by an option that is not exercised.[5] The benefits associated with the $(1 - \theta)R$ unexercised options will then be equal to $\omega(1 - \theta)R$, and this is the amount by which the consumer's surplus associated with the q^o calls will understate the consumer's willingness-to-pay for access.

Option demand springs from uncertainty, but uncertainty takes different forms. Many calls are contingent upon objective states of nature, while others are contingent upon subjective states of mind. A call to the fire department when lightning ignites a fire illustrates the former, while a call to a friend on a spur-of-the-moment illustrates the latter. The distinction is of importance because the consumer's preferences in the first case can be viewed as known and fixed, but random and therefore unknown in the second case.

Uncertain states of nature can be interpreted as risk (in the economist's sense of the word), so that assuming that the consumer is risk-averse, having access to the telephone system can be viewed as the purchase of an insurance policy. However, uncertain preferences are another matter, and it seems best to treat this form of uncertainty as the type of uncertainty defined many years ago by Frank Knight. With Knightian uncertainty, the various states of mind cannot be described by a probability distribution. This form of uncertainty is no less important than risk in giving rise to an option demand for telephone calls, but it is obviously

[4] I am ignoring the complications introduced by the access externality and the benefits associated with incoming calls.

[5] Since the value of this benefit will probably vary with the called involved, ω should be viewed as a mean.

much more difficult to deal with analytically. However, this need not concern us now, because there is little prospect at present of being able to distinguish empirically between the option demand arising from contingent states of nature and the option demand arising from uncertain preferences.[6]

At this point, we are left with the option demand associated with incoming calls. However, since incoming calls are usually beyond the control of the party receiving the call, it is reasonable to identify the option value for incoming calls with the benefits yielded by these calls. This done, there is nothing further to discuss since these option values will be already reflected in the benefits attributed to these calls.

The Opportunity Cost of Time

To this point, it has been assumed that the only constraint on a consumer's behavior is the income that is available to be spent. The assumption is that the consumer allocates available income between telephone calls and other goods and services in such a way as to maximize the utility that can be obtained. However, time is also a constraint on a consumer's behavior, for consumption does not occur instantaneously. A consumer must not only have the income to purchase a good, but also the time to consume it. Thus, proper analysis of consumer behavior must

[6] I do not wish to make too much of the randomness of preferences in this context, for what I feel is really the case is that calling behavior is random in the small rather than in the large. I do believe that preferences are stable in the sense that at the time that the purchase of access is being considered, consumers recognize that some calls will be subject to whim and fancy and plan accordingly. All of us have made calls, without subsequent regret, that were prompted by an ephemeral mood. That such occasions are likely to occur gives rise to an additional willingness to pay in order to have access to the telephone system.

treat both time and income as constraints.[7]

However, there is an important difference between time and income that is absolutely critical to the analysis. The amount of income that is available can be varied, but this is not the case with time. Nothing can change the fact that there are only 24 hours in a day. A consumer has to decide how to allocate these 24 hours between the time spent on the job earning income and the time spent in the home "consuming" this income and keeping the body and mind in order. Income, in contrast, while fixed in the short run, can be increased in the long run because of increased productivity and the fact that time can be reallocated between time spent on the job and time spent in the home.

In general, the consumer will allocate his time in such a way that the benefits from its use are equated on the margin. Doing this will maximize the amount of satisfaction that can be obtained from the time that is available. As a consumer's market wage increases, the value of time spent in earning money income increases, and if satisfaction is to continue to be maximized, adjustments must be made in order to bring the value of time spent in various activities into equality on the margin. There are a number of forms that these adjustments can take. As the wage increases, the same amount of income can now be earned with a reduced amount of labor time, and the consumer could take the increased labor productivity, not only in the form of increased money income, but also through increased leisure (including time spent in home production).

[7] See Becker (1965), Linder (1970), and Gronau (1970).

Alternatively, the consumer might continue the same number of hours on the job (or even increase these hours) and reallocate his consumption expenditures in such a way as to increase the productivity of the time spent in home consumption and production. What would occur in this case is that goods, which are now relatively less expensive in terms of time, will be substituted for time in home consumption and production.

The reason that time and the telephone are so closely related is that the telephone is a major vehicle for increasing the efficiency of time. As the market wage rate increases, the opportunity cost of time increases, and there is accordingly increased incentive to economize on its use. The telephone provides a means for doing so. As the Yellow Pages say, "Let your fingers do the walking." However, in delving into the relationship between time and the telephone, one has to look at the substitutes for the telephone, which at present essentially consist of the mail and travel (or telex and telegram in the case of overseas communication). The mail requires relatively little out-of-pocket expense, while travel generally requires a lot. Both are highly time intensive relative to the telephone. As a consequence, whenever the opportunity cost of time increases, consumers have an incentive to substitute the use of the telephone for mail and travel.

However, while the telephone is highly efficient in the use of time, a telephone call also requires time as an input. Indeed, recording machines aside, a telephone call requires the undivided attention of at least two parties. Consequently, the full cost of a call includes not only the out-of-pocket cost but also the opportunity cost of the time that goes into the call. In the case of a local call under flat-rate pricing for basic service, for example, the out-of-pocket cost of the call is zero. The actual cost of the call, however, is the opportunity cost of the time required for the call to be completed.[8]

II. A CRITIQUE OF THE EXISTING ECONOMETRIC LITERATURE ON TELECOMMUNICATIONS DEMAND[9]

The preceding section discussed the attributes of the telephone system that set telephone demand apart from the demand for most goods and services. These, to repeat, include (a) the distinction between access to the telephone network and use of the network once access has been acquired, (b) the presence of access and call externalities, (c) option demand as an important component of access demand, and (d) the opportunity cost of time. These attributes will provide a useful background in this section in reviewing the econometric literature on telecommunications demand. The literature on telecommunications demand is large and diverse, and the discussion here is highly abbreviated. Readers interested in a much more detailed review and critique are referred to chapters 3 and 4 of my monograph.

As mentioned, the distinction between access and use should provide the main point of departure in building models of telephone demand. This

[8] The opportunity cost of time has been virtually ignored in the empirical literature on telephone demand. Beauvais (1977) is the only study that I am aware of that attempts to include the cost of time in a meaningful way. [For a discussion of Beauvais' analysis, see Chapter 3 of Taylor (1980)]. Some indirect evidence of the importance of the opportunity cost of time is offered in several econometric models of intrastate toll demand that I have seen in which both the number of toll calls and the average duration of a call are explained as a function of income, price, and other variables. Calls are a positive function of income, but duration is a negative function, which is to say that as income increases the number of toll calls increases, but the average duration of a call decreases.

[9] This section is taken mostly from Chapter 5 of my monograph.

means that the analysis should be approached in stages, with the first stage focusing on the demand for access. The most meaningful quantity to explain in this regard is the number of main-station telephones for residential subscribers and the number of main stations plus PBX (private branch exchanges) extensions for business customers. Stage two of the analysis will then focus on the demand for use. However, the demand for use may itself need to be approached in stages, depending upon how use is priced. On the one hand, if a call is priced on a two-part tariff, in which the price for the initial period differs from the price of an overtime period (as is the case for toll call in the U.S. and Canada), then the demand for use should involve two equations -- one that explains the number of calls and one that explains average duration. On the other hand, if a call is not priced on a two-part tariff (as is the case in Sweden and the U.K.), a single equation that explains the number of conversation-minutes will suffice.

The access/use distinction is found in studies throughout the econometric literature, but it is center stage in only three, namely, Alleman (1977), Pousette (1976), and Waverman (1974). Alleman (whose focus is the U.S.), restricts his analysis to the demand for basic service, (i.e., access), but Pousette and Waverman estimate complete models, which is to say that they estimate equations for use, as well as for access. Unfortunately, the other studies that estimate equations for the number of telephones contain little by way of theoretical motivation. In most cases, the analysis is guided by the general principles of demand theory, but the access/use distinction as a peculiar feature of telephone demand is largely ignored. How system size is tied to use, and how use, in turn, depends upon system size is in general not considered.

Let me illustrate the importance of the access/use distinction with Waverman's model for Sweden. Waverman, to recall, estimates equations for the number of main stations, local use, and toll use. The equations for main stations and local use will illustrate the point that I wish to make. Income is a predictor in both of the equations, the price of local use is a predictor in the equation for local use, but not in the equation for main stations, and the number of main stations is a predictor in the equation for local use. Thus, local use depends both directly and indirectly on income (since local use depends on the number of main stations, which, in turn depends on income), but only directly on the price of local use. (There is no indirect price effect via the number of main stations in Waverman's model, since main stations do not depend on the price of local use.) Suppose, now, that there is a change in income. What will be the impact on local use as measured (say) by the income elasticity of demand? Waverman's equation for local use is a Koyck logarithmic distributed-lag model, so that three income elasticities can be adduced, a "short-run" short-run elasticity and a "long-run" short-run elasticity, both of which are conditional on the number of main stations, and a long-run elasticity in which the number of main stations is allowed to vary in response to the change in income. Waverman's estimates of these three elasticities are 0.23, 0.32, and 1.25. Thus we see that in Waverman's model the indirect effect on local use of a change in income that arises through the equations for main stations is substantial.

The moral of the story is as follows: If the focus is entirely on estimating the effects on usage of changes in the price of usage, an access/use framework is not critical -- so long, that is, as usage is assumed to be conditional on the number of main stations -- as any impact of the price change on the number of main stations is probably small enough that it can be ignored. However, if the focus is on estimating the impact on usage of changes in income, an access/use framework *is* critical, because to ignore the indirect effect on usage that arises through the adjustment in main stations is likely to lead to a serious underestimate of the total effect of a change in income.[10]

Let me now turn to the access and call externalitites. The access externality can in principle be taken into account by making the demand for access depend on system size. One way that this can be done is to relate the current number of main stations to the number of main stations in the preceding period. If there are no other dynamics, a positive access externality will be reflected in a coefficient on the preceding period's main stations that is greater than 1. The call externality is

[10] The state-intrastate toll demand models in the U.S., of which between 30 and 35 existed in 1978, come to mind at this point. Most of the state models assume toll demand to be dependent on the number of main stations, but the models do not include equations for the number of main stations. As just noted, this does not create any problems so long as the focus is on price changes and there exists good exogenous forecasts of the number of main stations. However, there will probably be occasions when the models will be used to forecast the impact of changes in income, and care must be taken to allow for feedbacks on the number of main stations. The explicit use of an access/use framework automatically does this.

more difficult to measure empirically, although a case can be made that it too can be represented by system size. The basis for this is discussed in my monograph.

The evidence regarding the access and use externalities is thus to be found in the equations that include a measure of system size as a predictor. This occurs in many of the state intrastate toll demand models and in the models of Feldman (1976), Davis *et al.* (1973), Pousette (1976), and Waverman (1974). In most of these models, system size is measured by the number of telephones less residence extensions, but in a few of the state models, the number of households or the population is used as a surrogate. Most of the equations in question refer to the demand for use. The only ones that involve the demand for access are the equation of Davis *et al.* for the total number of telephones (less residence extensions) and Waverman's equation for the number of main stations in Canada.

Generally speaking, the evidence concerning the two consumption externalities is inconclusive. The strongest suggestion that the externalities may be of some importance is given in Waverman's equation for local use in Sweden, which has an elasticity with respect to the number of telephones of 1.19.[11] However, the fact that the existing

[11] Additional evidence concerning the externalities is found in the studies of Infosino (1976) and Wang (1976). Infosino finds that the number of local calls per line in a sample of residential customers in Los Angeles and San Francisco is positively related to the telephone density of the exchange, while Wang finds that the demand elasticity for yellow-page advertisements of a given size with respect to system size is greater than 1. Wang's result implies that, with the size of an advertisement and price held constant, space demand is stronger in a larger system.

evidence is weak and mixed is hardly surprising since the empirical literature has not explicitly focused on the consumption externalities as factors to be taken into account. Many of the state toll demand models, for example, assume them away *a priori* by defining the dependent variable as the number of messages per main station (or price-deflated revenues per main station). Moreover, it is not the case that the externalities have been considered and then dismissed as unimportant, for in general they have simply been ignored. Waverman, for example, does not see in his results for Sweden the suggestion that the externalities may be consequential, and Pousette, in his otherwise admirable study, ignores them altogether.

Let me now move on to option demand. The literature does not provide any empirical evidence, even inadvertently, regarding its importance. However, this too, is hardly surprising, for while option demand is an appealing concept, it is not easily given to measurement. In my monograph, it is suggested that option demand might be expected to be relatively more important in rural exchanges than in urban exchanges, and if so, the access-demand elasticity with respect to the price of access should be smaller in rural exchanges. This might be tested with the data set constructed from the 1970 U.S. Census used by Perl (1978). Also, it seems that option demand might be a factor in many subscribers' apparent preference for flat-rate pricing of local service over measured service. While it is not clear how this idea can be tested empirically, it has important implications for the pricing of access, and is therefore worthy of attention.

Let me now turn to some other matters. A major deficiency in the empirical literature is the treatment of prices. Most telephone services are priced on a multi-part tariff, and this has a number of important implications. With a multi-part tariff, one has to distinguish between the marginal price and the intra-marginal prices; moreover, in some cases, a multi-part tariff changes the basic logic of the demand model. A toll call in the U.S. and Canada, for example, is priced on a two-part tariff, since the price per minute is less for overtime periods than for the initial period. If the goal is to explain the number of conversation-minutes, the appropriate procedure (as discussed earlier in this section) is to decompose conversation-minutes into the number of calls and the average duration of a call. The number of calls should then depend (besides income, etc) primarily on the price of the initial period, while the average duration of a call should depend primarily on the price of an overtime period.

The principles that underly these conclusions are discussed in detail in my monograph, but it is useful to summarize them here.[12] When a good is priced on a multi-part tariff, the separate components of the tariff affect a consumer's behavior in different ways. In equilibrium, the consumer equates marginal rates of substitution between pairs of goods to the ratios of their respective *marginal* prices. However, when the good involved is defined in several dimensions, then a tariff that is marginal in one dimension may be intra-marginal or extra-marginal in another

[12] See also Taylor, Blattenberger, and Rennhack (1981).

dimension. With a toll call, there are two decisions to be made: whether to make the call, and how long to talk. The price of the initial period is the price that is most relevant to whether to make the call, the time-of-day to make the call, and whether to direct dial, while the price of an overtime period is the price that is most relevant to how long to speak.

In general, these considerations are not reflected in the empirical literature. There are only two studies that seem genuinely cognizant of the problems that multi-part tariffs pose, namely, Deschamps (1974) and Pousette (1976). Deschamps discusses explicitly the complications created by multi-part tariffs, and his study stands out in this regard. Pousette's actions, on the other hand, speak louder than his words: although he does not verbally distinguish between the price of access and the price of use, his equations for new connections contain a price index that is a weighted average of the subscription fee and the call charge, while the equations for use contain only the call charge. Combining the access and use charges into a single index is not the ideal procedure, but it is certainly a step in the right direction.[13]

[13] The studies of Waverman (1974), Perl (1978), and Larsen and McCleary (1970) should also be mentioned in this regard. Waverman distinguishes between the price of access and the price of use in his equations for Sweden, but not for Canada and the U.K., and Perl, in his analysis of access demand for the U.S., treats the service-connection charge separately from the monthly service charge and also distinguishes between exchanges with measured local service and exchanges with flat-rate service. Finally, Larsen and McCleary, in their analysis of toll traffic between pairs of U.S. states, examine both the average charge per call and the average overtime charge per call. Unfortunately, though, Larsen and McCleary calculate both measures of price from *ex post* data.

The failure to distinguish among the various components of a multipart tariff usually leads to the use of an average price, and the worst situation is when the average price is obtained by dividing revenues by the quantity that is being explained. An average price for toll calls, for example, is frequently obtained by dividing toll revenues by the number of toll calls. Such an *ex post* procedure is to be avoided because it necessarily establishes a negative relationship between quantity demanded and price.[14] Perhaps the most serious lapse in this connection in the telephone demand literature is by Beauvais (1977) in his study of the demand for local calls. Beauvais defines an average price by dividing the monthly service charge by the number of local calls. However, this creates a very serious bias in the estimate of the price elasticity because most of the variation in the price variable is caused by the variation across subscribers in the number of local calls. Other studies in which price is calculated as an *ex post* average price include Feldman (1976), Kwok, Lee, and Pearce (1975), Larsen and McCleary (1970), and Rash (1972).

The state toll demand models mentioned in footnote 10 all use a price index for intrastate MTS calls for the price variable. Laspeyres indices are used in about two-thirds of the models, while chain-weighted indices are used in the rest. Never do the weights depend on current-period quantities demanded, so that the price variables used in these models do not

[14] This is a long-standing problem in the analysis of electricity demand. See Taylor (1975).

suffer from simultaneous-equations bias. However, the price variables defined in this way do have a problem in that the initial-period and overtime-period charges are combined in a single index. This is appropriate for the models in which the dependent variable is price-deflated revenues, in which case the dependent variable is a measure of conversation-minutes, but is is not appropriate for the models in which the dependent variable is the number of messages. In this case, the price of the initial period should be separated from the price of an overtime period.

Another shortcoming in the empirical literature is a focus on the number of calls, as opposed to conversation-minutes. As has been noted, when the price of a toll call consists of an initial charge plus a charge that depends on the number of overtime periods, one should explain duration as well as the number of messages. If the price of a call were independent of duration, the latter could be ignored, but this is usually not the case, so that the possible dependence of duration on price must be taken into account. Only four studies in the literature do this, Feldman (1976), Gale (1974), Pousette (1976), and Waverman (1974). The studies of Feldman and Gale provide the most detailed analyses and, in both studies, duration is found to be negatively related to price.[15] Pousette and Waverman both explain the number of "pulses" (which is a physical measure of holding time used in most Western European countries), and duration *per se* is not singled out for analysis. However, Waverman also estimates an equation for the number of messages, and since the price

[15] Gale's study is unique in that it is the only study that focuses exclusively on the dependence of duration on price.

elasticity in this equation is smaller than in his pulse equation, one can infer that there is also a nonzero price elasticitv for duration.

The ignoring of duration is especially apparent in the U.S. state toll-demand models. None of the state models have equations that focus directly on duration. About half of the state models whose elasticities are tabulated in Chapter 4 of my monograph have the number of messages as the dependent variable, while the dependent variable in the remaining state models is price-deflated revenues. Since toll revenues can be decomposed into the product of the number of calls and the average revenue from a call, price-deflated revenues in principle correspond to the product of the number of calls and their average duration. The impact of a price change on average duration is thus reflected in the estimated price elasticities.

Whether the dependent variable in the U.S. state models should be the number of messages or price-deflated revenues, has been a subject of some debate. The state models have been developed mainly as planning tools and for isolating market reactions to tariff changes in rate filings before state Public Utility Commissions. In most cases, their primary use in rate filings has been to estimate the impact on toll revenues of a nonzero price elasticity of demand. For the models with the number of messages as the dependent variable, revenues (after repression[16]) are calculated by multiplying the estimated number of messages by a

[16] Repression in this context refers to the impact on revenues of a rate change when there is a nonzero price elasticity of demand.

repriced average revenue per message. For the models with price-deflated revenues as the dependent variable, after-repression revenues are calculated by multiplying the estimate of repression-adjusted real revenue by a price incex which reflects the new rates.

Neither of these procedures is ideal, the reason being that the number of messages and the average duration of a message respond in different ways to changes in the tariff structure. In the models with price-deflated revenues as the dependent variable, the price of a call is represented by a price index which combines the charge for the initial period with the charge for an overtime period, which means that the responses of messages and average duration are reflected in a single elasticity. In situations where the structure of the tariff schedule is to be changed, not allowing messages and duration to respond differently could lead to serious forecasting errors. However, the problem in these models is not so much in the use of price-deflated revenues as the dependent variable, but rather in the use of a single price index. The use of two price indices, one for the initial-period charge and the other for the overtime-period charge, would overcome the problem. In contrast, the problem in the models with the number of messages as the dependent variable is that an equation is missing, namely, an equation for average duration. Either the models do not make any allowance at all for the effect of a rate change on duration or else it is assumed that the price elasticity for messages also applies to duration. Neither of these procedures is satisfactory, but the solution (at least in principle) is readily apparent: estimate an equation for average duration.

Let me now turn to some questions of dynamics. The models using the access/use distinction are dynamic by definition -- since use is predicated on system size -- and distributed-lag models are used extensively, especially in the analyses of U.S. intrastate toll demand. The tables in Chapters 3 and 4 of my monograph containing existing estimates of price and income elasticities of demand provide a great deal of evidence that telephone demand is indeed a dynamic phenomenon, for when dynamic models are specified, the estimated long-run elasticities are nearly always considerably larger than the short-run elasticities.

Perhaps the biggest problem connected with the treatment of dynamics in the empirical literature is that distributed-lag models are frequently forced to do too much. Dynamic adjustment can arise from two sources: the first reflects the dynamics inherent in the consumption of a service generated by a complementary durable good,[17] while the second reflects inertia that may exist in the short run.[18] A distributed-lag model can capture both types of dynamic processes, but they cannot be sparately identified in the same model. Usually, this cannot be avoided, since estimates of the stock of complementary durable goods either do not exist

[17] In the present context, the durable good is the telephone system, while the service generated by the durable good is the use of the system.

[18] To dispel possible confusion, let me be more specific. The dynamics inherent in the access/use division represent the traditional distinction between the short run and the long run in the presence of a durable good. Suppose that the telephone system is in steady-state equilibrium (ignore the complications introduced by the access externality), and let this equilibrium be distributed by an increase (say) in income. In the short run, there will be an adjustment in the number of calls that are made using the existing stock of telephones (and possibly also in average duration). In the long run, the stock of telephones may also adjust. However, it may be that in the short run (when the stock of telephones is fixed) there is a delay in adjusting the number of calls that are made to the higher income.

or else are of too poor quality to be used. But in the telephone industry, this is not the case, for the data on the stock of telephones are in general quite good. The access/use distinction can accordingly be modeled directly (thereby taking into account the dynamics associated with the consumption of a service generated by a durable good), while a distributed-lag model can then focus exclusively on capturing short-run inertia.

On the other hand, one can question whether the Koyck model, which has been used in the vast majority of cases, may be too restrictive. As is well-known, the Koyck model postulates geometric decay in the distributed lag, and also constrains each independent variable to have the same lag structure. Frequently, these restrictions are probably unrealistic, but are resorted to because multicollinearity precludes meaningful estimation of separate lag structures. However, other more flexible, distributed-lag models exist, and these should be analyzed with the purpose of seeing whether price and income have different lag structures. The Almon polynomial-lag model would be a convenient (but not the only) model to use to this end. Transfer functions, which have been used in a few studies with promising results, also merit consideration, particularly when quarterly or monthly data are being analyzed.[19]

Another thing which has not been adequately explored in the empirical literature is the time-series/cross-section nature of much of the telephone

[19] For a discussion of transfer functions, see Box and Jenkins (1976). Cf. also Fask and Robinson (1977).

data base.[20] The data collected in AT & T's CMDS data base would allow for this, and so, too, would the data collected at the state level by the operating telephone companies. The major benefit from pooling is the increased variation in the independent variables. However, care must be taken that the structures being pooled are homogeneous -- i.e., that regression coefficients are constant across observational units, and similarly for the structure of the error term. Traditionally, the covariance model has been the model that has been employed in pooled time-series/cross-section analysis, but recent years have seen an increasing use of the variance-components model. Random coefficients models may also have a role to play, especially in situations where price or income elasticities are found to vary regionally.[21]

[20] There are only six studies in the literature that estimate models using pooled time-series/cross-section data. Kearns (1978) and Reitman (1977) pool time-series data across states in estimating models of the demand for residence extensions (Kearns, Reitman) and the demand for total vertical services (Reitman). Stuntebeck (1976) and Wert (1976) both used pooled data in analyzing the daytime/nighttime composition of toll traffic. Finally, Deschamps (1974) uses a pooled time-series/cross-section data set in analyzing toll demand in Belgium, and Rea and Lage (1978) do the same for international telephone, telex, and telegraph demand.

[21] Of the studies just listed, Deschamps and Rea and Lage are the only ones to use a variance-components framework. Kearns, Stuntebeck, and Wert use a covariance model, but Reitman pools directly without either the covariance or variance-components adjustment. The consequences of pooling directly are especially severe in Reitman's study, for a model is used in which the lagged value of the dependent variable is included as a predictor. For models of this type, a covariance or variance-components framework is compelling, otherwise, the individual state effects will be reflected in the lagged value of the dependent variable, and the estimate of its coefficient can be severely biased.

Since many of the econometric models of telephone demand are used in rate filings, the basic canons of econometrics need to be given a great deal of attention. Otherwise, the results of the models, or even the models themselves, may be challenged on the grounds that proper econometric and statistical procedures have not been followed. Most of the time, econometricians take many things for granted, especially when communicating with other econometricians; to do otherwise would be tedious, repetitive, and time-consuming. The validity of the t- and F-tests, for example, requires that the error term be normally distributed, but rarely, is normality tested for explicitly. Usually, the Classical Central Limit Theorem is relied upon to provide normality, but failing this, the econometricians know that the t- and F-tests are robust in the face of even quite substantial departures from normality. However, the people that ultimately have to be convinced in rate filings are not other econometricians, but rather those who may have little understanding or appreciation of econometric procedures. As a consequence, laxity and possible errors in procedure can be made to seem much more important than they in fact are. Hence, in using a model in a rate filing, formal procedures need to be followed with especial care, including (to return to the example) the explicit test for normality in the error term.[22]

[22] I have singled out normality of the error term because it is not usually considered a trouble point. The conventional list of statistical problems includes autocorrelation, heteroscedasticity, and multicollinearity. Autocorrelation is nearly always a problem with aggregate time-series data, and extreme care should be exercised in checking for autocorrelation in situations where t-ratios are around 2. Heteroscedasticity, on the other hand, is more likely to be a problem with cross-section data than with time-series data. In general, there is probably more laxity in checking for heteroscedasticity than in checking for autocorrelation. Finally, multicollinearity is nearly always present in some degree, no matter what the source of the data, although it is typically strongest with time-series data.

Common sense suggests that access to the telephone system should be more of a necessity than local use and that local use should be more of a necessity than toll use. This would mean that the income elasticity for access would be smaller than the income elasticity for local use, which, in turn, would be smaller than the income elasticity for toll use. In general, we should expect the same relationships to hold among the price elasticities, although the reasoning is a bit more subtle. The effect on demand of a change in price, to recall, consists of two terms, an income effect and a substitution effect. If the substitution effect is held constant, the price elasticities will in general increase *pari passu* with the income elasticities. However, common sense also suggests that the substitution effect should be very weak for access, and weaker for local use than for toll use. Thus, both the income effect and the substitution effect should make for progressively larger price elasticities as we move from access to local use, to short-haul toll, etc.

The empirical results tabulated in Chapters 3 and 4 of my monograph support these views. In general, the estimated elasticities for access are smaller than the elasticities for local use, which are smaller than the elasticities for toll demand. This is true for both the income and the price elasticities. Moreover, the empirical results also indicate that the elasticities for toll demand vary with distance, being smaller for short-haul than for long-haul calls. The elasticities in the U.S. intrastate models are, in general smaller than in the interstate models.

In Table 1, I have tabulated some point estimates from my monograph of price and income elasticities for the demand for access, local calls, and long-distance calls and for the duration of long-distance calls. It should be emphasized that these estimates are my own interpretation of the existing empirical record. They are accordingly highly subjective, and are based on evidence from other countries as well as from the U.S. While the estimates tend to be near the midpoint of existing estimates, I have implicitly given some studies more weight than others. The estimates refer to steady-state, long-run elasticities, and are for residence and business demands combined, for the empirical record at this stage will not support an attempt to distinguish between residence and business customers. As a measure of the uncertainty to be associated with the estimates, I have appended a range to each estimate. These ranges are also subjective, and reflect my own views as to the intervals within which the true elasticities are likely to lie.

The greatest uncertainty attaches to the estimates of the income elasticities of the demand for use, particularly for local calls and interstate toll calls. I think that one can conclude that the income elasticity for toll calls is greater than 1, but how much greater is still an open question. Considerable uncertainty also surrounds the price elasticity for interstate toll, especially with respect to the critical value of 1 (in absolute value). There is a strong feeling within the telephone industry that the price elasticity for interstate toll is less than 1, and, indeed, is most likely in the neighborhood of 0.5. My own view at this juncture is that the long-run price elasticity for interstate toll in general is less than 1, but that the value for long-haul calls may be closer to 1 than to 0.5.

Table 1

POINT AND INTERVAL ESTIMATES
OF PRICE AND INCOME ELASTICITIES
OF DEMAND FOR SELECTED TELEPHONE SERVICES

ELASTICITY

Type of Demand	Service-Connection Charge	Monthly Service Charge	Toll Price	Income
Access	-0.03 (±0.01)	-0.10 (±0.09)	--	0.50 (±0.10)
Local Calls	--	-0.20 (±0.05)	--	1.00 (±0.40)
Toll Calls (conversation-minutes)				
Intrastate	--	--	-0.65 (±0.15)	1.25 (±0.25)
Interstate	--	--	-0.75 (±0.20)	1.50 (±0.40)
International Calls	--	--	-0.90 (±0.30)	1.70 (±0.40)
Duration of Toll Calls	--	--	-0.15 (±0.05)	0.25 (±0.10)

Source: These estimates refer to long-run, steady state elasticities. The estimates for toll calls refer to conversation-minutes, rather than to just the number of messages. The estimates reflect my own interpretation of the empirical record (for both foreign countries and the United States) and are thus highly subjective. The numbers in parentheses provide an interval within which I feel it is highly likely that the "true" elasticity lies.

The entries in the table provide only a partial listing of the categories of telephone demand. I have not included any elasticities for WATS and private line, vertical services, and coin stations, and, as mentioned, I have not attempted a residence/business breakdown. The empirical evidence in all of these areas is too weak for the tabulation of "best-guess" point estimates of elasticities. For WATS and private line, existing evidence suggests that own-price and income elasticities are more or less the same as for MTS, and there is solid support in the Feldman study (1976) that WATS, private line, and MTS are strong substitutes. Yet, interestingly, there is a suggestion in the study of Subissati (1973) that WATS and MTS (in Canada) may be complements. What Subissati finds is that current expenditures for MTS are positively related to the first difference in expenditures for WATS. Subissati suggests that this may reflect a stimulus to total toll calling induced by the presence of an additional way to make toll calls. If this is in fact the case, then WATS, private line, and MTS may actually be complements in the long run, but substitutes in the short run. Whatever, the tradeoff between and among WATS, private line, and MTS is an important area for future research.

Let me now turn to some implications of the findings concerning the price and income elasticities of demand. To begin with, it must be emphasized that price elasticities exist; which is to say that, contrary to the views of many, they are not zero. On the other hand, it does appear that, except possibly for very long-haul toll calls, telephone Demand is inelastic. The demand for access, in particular, appears to be very price inelastic. Also, there is some evidence that the "transient" component of inward and outward movement is quite sensitive to the level of the service-connection charge.

The fact that the price elasticity for local use is small means that the telephone companies can look to the local market as a place to recoup the revenues that are almost certain to be lost in the private line, toll, and terminal-equipment markets as a result of competition. A key question, therefore, is whether the price of access to residential customers is to continued to be subsidized. In the past, the subsidization of residential access, primarily from "contributions" generated in the toll market, has been justified in terms of fostering universal service. Residential rates have been kept artificially low (in terms of cost) in order to make telephone service available to essentially anyone who wanted it. However, the access externality is also an issue. If the access externality exists, then optimal social pricing requires that the price of access continue to be subsidized.[23] If the externality does not exist or is quantitatively unimportant, then optimal social pricing would require that the price of access be set equal to marginal cost.[24] Thus, we are once again reminded of the importance of establishing whether the access

[23] By optimal social pricing in this context, I mean a policy that has the objective of maximizing the sum of consumers' and producers' surplus. See Zajac (1979) and Wenders (1980).

[24] Because of rate-of-return (or other regulatory) constraints, optimal social pricing might require that the price for access deviate from marginal cost in a way that depends upon the price elasticity of the demand for access. This is usually referred to as Ramsey-pricing after F. P. Ramsey (1927). See Baumol and Bradford (1970) and Zajac (1979).

externality exists and estimating its quantitative magnitude.[25]

Let us suppose, for now, that the access externality is unimportant and that the telephone companies set out to increase the price of access to the level of marginal cost. What might be expected to happen? My own view is that there would probably be considerable resistance by residential customers. Besides just common sense, I base this observation on the empirical evidence offered by Perl (1978) which shows the dependence of the price elasticity for access on the level of income and the level of the access price. Perl's results show that the elasticity with respect to the access price decreases with the level of income and increases with the level of the monthly service charge. The access price elasticity is accordingly largest (although never absolutely large) for low-income households facing a high monthly service charge. On the other hand, the price elasticity is smallest (and very small indeed) for high-income households facing a low monthly service charge. One can, therefore, conclude that most of the access price elasticity arises from the income effect, rather than from the substitution effect. However, the income effect implies a loss in consumer welfare, whereas the substitution effect does not. Consequently, as the price of access is moved toward marginal cost, one should expect (probably very spirited) consumer resistance, particularly on the part of low-income households living in areas where the monthly service charge is already fairly high.

[25] If the access externality is quantitatively significant and if the telephone companies cannot subsidize access from the local-use market because of competition, how the subsidy is to be financed obviously becomes an important social question.

Thus far in the discussion, the focus has been mostly on price elasticities; income elasticities have been treated largely in passing. The reason for this is that for the most part, income elasticities are not controversial, but this is not the case with price elasticities, especially during a period of frequent (and sometimes large) rate increases. However, income elasticities are very important to the telephone industry because they indicate how, holding prices, technology, and other non-income determinants of demand constant, the industry will develop over time, which markets will require the most additional investment, and where additional revenues will accrue. In the markets where the aggregate income elasticity is greater than 1, the growth in revenues will be faster than the growth of the general economy, while in the markets where the aggregate income elasticity is less than 1, the growth in revenues will be slower than the growth of the general economy. Of course, non-income factors do not remain constant, so that revenue growth in individual markets can be quite different than that implied by income elasticities alone. However, in general, one should expect revenues to grow most rapidly in the markets with the highest income elasticities of demand.

As was noted earlier, the estimates of income elasticities are generally substantial, although there is a lot of variation in the estimates, especially in the ones for local use and interstate toll.[26] Table

[26] See Tables 4 and 5 in Chapter 3 of my monograph.

1 suggests point estimates of at least 1 in all of the major telephone markets except for access. As noted, however, a great deal of uncertainty surrounds the estimate for local use, so that this elasticity could very well be less than 1. The income elasticity for toll calls, on the other hand, is almost certainly greater than 1, particularly in the long-haul interstate market. And there seems little question but that the elasticity for overseas calls is substantially in excess of 1. Indeed, for a number of years overseas revenues have been the fastest growing component of total Bell System revenues. Already these revenues are making an important contribution to Bell System profits, and, if present trends continue, there will be a time when they are the most important contributor.

IV. CURRENT ISSUES AND PROBLEMS

In this section, I shall present some concluding observations regarding the present state of demand analysis in the telephone industry and some suggestions as to where we might go from here. On the whole, the quality of analysis in the empirical literature is good. The models that have been analyzed are essentially state-of-the-art for applied demand analysis, and this is also true for the econometric techniques that have been used in estimation. The empirical literature contains a number of good studies and several really excellent ones. Included among the latter are Deschamps (1974), Gale (1974), Griffin (1981), Feldman (1976), Irish (1974), Larsen and McCleary (1970), Mahan (1980), Pavarini (1975, 1976, 1979), Perl (1978), Pousette (1976), Stuntebeck (1976), and Waverman (1974). The theoretical literature also contains some first-rate contributions, with the list headed by Artle and Averous (1973), Littlechild (1975), Rohlfs (1974), and Squire (1973).[27]

Still, the quality of the empirical literature falls short of where it ought to be. The biggest problem is that the empirical and theoretical analyses of telephone demand have been like two ships passing in the night. The best empirical work has ignored the best theoretical work, and *vice versa.* Clearly, the theoreticians and the applied analysts need to join forces. Telephone demand modelers would also benefit from greater contact with the experience of demand analysts in other areas, particularly energy demand. Pooled time-series/cross-section models have been used extensively in analyzing energy demand, and the experience that has accumulated there is clearly relevant to the increased use of similar models for telephone demand. Energy demand analysts have also had extensive empirical experience in dealing with the problems caused by multi-part tariffs.[28]

It was noted earlier that existing estimates of price elasticities of demand solidly support the conclusion that telephone price elasticities are different from zero. However, a great deal of uncertainty surrounds virtually all of the estimates, and one of the major tasks of

[27] I view the studies cited here as constituting a "bare-bones" reading list for anyone wishing to become familiar with the literature on telephone demand. Brandon (1981) and Mitchell (1978) are also highly recommended, as are also two older studies, Kraepelien (1958) and Leunbach (1958).

[28] See, for example, Taylor, Blattenberger, and Rennhack (1981) and Acton, Mitchell, and Mowill (1976).

future research is to reduce the zones of uncertainty. There is some evidence that the price elasticity for long-haul toll calls may be as large as -1, and since this is a decidedly critical value, it is particularly important that the uncertainty associated with it be reduced. The efforts to do this, however, should distinguish more carefully than in the past between the number of calls and the duration of calls, and price indices should be used that capture the essential characteristics of multi-part tariffs. There is also considerable uncertainty surrounding the estimates of income elasticities, especially for local and long-haul toll calls. The existing estimates do suggest, though, that the income elasticities are at least 1 in all of the major markets except for access, and substantially greater than 1 in the long-haul toll and international markets. In short, telecommunications should continue to be a growth industry.

When we look at the empirical literature as a whole, access and local use have received relatively little attention in comparison with toll, and this should change. The demand for terminal equipment is also under-researched, as are also WATS and private line. Indeed, WATS and private line (particularly private line) have been virtually ignored. International demand has received some attention, but not nearly as much as its rapid growth warrants. Finally, business demand has received scarcely any attention in relation to residential demand, and this too should be corrected.

That toll demand has been the center of attention is readily understandable, for rate activity in recent years has tended to concentrate on the toll markets and toll is relatively easy to model. However, in view of the rather radical changes now taking place in the telecommunications industry, the research focus needs to change, and much greater attention should be given to access and local use. As competition in the private-line, toll, and terminal-equipment markets creates the pressure for additional revenues from access and local use, the industry and its regulators need much better information than now exists on how customers -- especially residential customers -- will react to higher access charges. A closely related question is how residential customers will react to the paced conversion to measured local service that is increasingly becoming the industry policy.[29] A major research effort is currently in progress at Bell Laboratories to examine these and related questions, but data are scarce and progress is slow and expensive.

Concerning toll demand, we have already mentioned the need to reduce the uncertainty blanketing the long-haul toll price elasticity. Also, now that competition is a factor in many of the intercity markets, knowledge of point-to-point price elasticities is of obvious interest. One of the biggest unanswered questions in toll concerns the tradeoffs among private line, WATS, and MTS for business customers, but the estimation of these tradeoffs needs to be approached in a well-articulated

[29] For discussion of measured local service from the Bell System perspective, see the recent articles in the *Public Utilities Fortnightly* by Garfinkel and Linhart (1979, 1980) and Cosgrove and Linhart (1979). See also Wenders (1981).

model of business demand. Some recent work by McFadden and Train (1979) provides some interesting and useful new suggestions in this direction.

DEMANDE ET CONSOMMATION TELEPHONIQUES: UN MODELE RESIDENTIEL GLOBAL

N. CURIEN

E. VILMIN

Direction Générale des Télécommunications

P.T.T. France

INTRODUCTION ET RESUME

Dans la lignée de la littérature sur la demande téléphonique([1] à [4]), on présente au § 1 un modèle microéconomique "intégré" permettant d'expliquer conjointement, en milieu résidentiel, la décision d'abonnement et le niveau de consommation téléphonique. On reprend la logique habituelle de maximisation d'une fonction d'utilité sous contrainte de revenu ; cependant des hypothèses vraisemblables sur la forme de cette utilité, déjà présentées dans un travail précédent [6] permettent d'expliciter les fonctions de revenu seuil d'accès et de niveau de consommation sans recourir à l'approximation du surplus. En procédant par agrégation sur les individus, le modèle est ensuite développé dans les deux directions : étude de la demande de raccordement (§ 2) et étude de la consommation moyenne par ligne (§ 3). La logique des décisions individuelles, telle que décrite par le modèle microéconomique générateur, permet de classer les variables explicatives de la demande et de la consommation (revenus, tarifs, ancienneté d'abonnement, externalité de l'offre), puis d'établir des conjectures théoriques sur les interrelations entre les évolutions de ces deux quantités : en particulier, l'effet négatif produit par la croissance du parc sur la croissance de la consommation. Les ajustements économétriques, présentés et discutés en détail, valident une grande partie les conjectures théoriques ; une attention particulière est portée aux élasticités aux tarifs.

INTRODUCTION AND ABSTRACT

In the framework of the theory of telephone demand ([1] à [4]), a microeconomic integrated model is proposed in section 1, taking into account both residential demand for access and demand for use. The classical utility maximization under budget constraint is carried out and some assumptions on the utility function (as introduced in a previous work [6]) allow to explicitly derive the thresehold income for access and the level of use, without using the surplus approximation. By aggregating, modelization is then develcpped in both directions : access demand (section 2) and traffic demand per main-station (section 3). The causality of individual decisions, as described in section 1 by the microeconomic generating model, first leads to identifying and classifying variables affecting access and traffic, as income, tariffs, supply externality, habit of use. Theory then provides conjectures about the relationship between the two aspects of demand as, for instance, the negative impact on usage produced by the growth of the penetration rate. Econometric estimations are presented and discussed in details and they give good support to theoritical predictions. Price elasticities are the subject of a particular attention.

1. LE MODELE MICROECONOMIQUE DE DECISION

1.1. La fonction d'utilité

On se place dans le cadre classique de la théorie microéconomique de la consommation et, au niveau d'un consommateur individuel i, on désigne par :

. x la quantité consommée en téléphone sur une période(exprimée en taxes de base)

. $\mathcal{T}$ le taux de pénétration téléphonique réel (instances non comprises), traduisant le niveau de l'offre

. T_i l'ancienneté d'abonnement téléphonique ($T_i = 0$ si i n'est pas abonné)

. X la quantité consommée sur une période de tous les biens non téléphoniques, supposés rassemblés dans un agrégat unique.

Nous admettrons que l'utilité en x et X du consommateur i peut être représentée par une fonction décomposable du type :

(1) $U_i(x,X) = Xu_i(x,\mathcal{T},T_i)$

où la composante téléphonique $u_i(x,\mathcal{T},T_i)$ est :

. positive, croissante, concave en x ($u_i > 0$, $\frac{\delta u_i}{\delta x} > 0$, $\frac{\delta^2 u_i}{\delta x^2} < 0$)

. croissante en $\mathcal{T}$: $\frac{\delta u}{\delta \mathcal{T}} > 0$

. telle que $u(0, ., .) \neq 0$, $u(\infty , ., .) \neq \infty$

Une fonction d'utilité telle que (1) traduit plusieurs aspects :

i) les propriétés classiques du préordre de préférence sur l'espace de consommation $\{ x \geqslant 0, X \geqslant 0 \}$: continuité, croissance, convexité (cf. par exemple [5])

ii) la marginalité et l'isolabilité du bien téléphonique "x" par rapport à l'ensemble des biens "X" (voir à ce sujet [6] et [7])

iii) l'externalité positive de l'offre téléphonique, représentée par le taux d'équipement $\mathcal{T}$ (cf [4]).

iiii) l'influence de la durée de consommation T_i du bien "x" sur la satisfaction liée à cette consommation.

La dépendance en T_i de l'utilité téléphonique u_in'est pas classique et mérite justification : elle permet en premier lieu de prendre en compte un éventuel glissement d'utilité lorsque le consommateur passe de l'état non abonné ($T_i=0$) à l'état abonné ($T_i > 0$) ; en second lieu, elle est confortée par les données empiriques qui montrent (cf § 3.2) que la dynamique d'évolution de la consommation est différenciée selon la valeur de T_i, les abonnés récents ayant un trend de consommation supérieur à celui des anciens.

1.2. Décisions de raccordement et de consommation téléphonique

Pour accéder à la consommation x au prix marginal p (valeur de la taxe de base), le consommateur doit acquitter une charge fixe A : en régime permanent de consommation ($T_i > 0$), cette charge correspond à la taxe d'abonnement ; au moment de la décision initiale de raccordement ($T_i = 0$) elle est majorée d'un amortissement de la taxe de raccordement.

La période de décision initiale de raccordement ($T_i=0$) et les périodes ultérieures de consommation ($T_i > 0$) peuvent être décrites selon un même schéma d'analyse, l'usager devant à chaque période effectuer deux choix :

- la décision d'accès ($T_i=0$) ou de maintien d'accès ($T_i > 0$)
- la détermination de sa consommation s'il décide d'accéder ($T_i=0$) ou de maintenir l'accès ($T_i > 0$).

La formalisation est la suivante : l'usager i, de revenu R_i, confronté au niveau général des prix P, réalise d'abord le programme d'optimisation de son utilité, conditionnel à l'accès :

$$(2) \quad \begin{cases} \text{Max } X u_i(x, \tau, T_i) \\ px + PX = R_i - A \end{cases}$$

puis le programme conditionnel au non-accès :

$$(3) \quad \begin{cases} \text{Max } X u_i(x, \tau, T_i) \\ x = 0 \\ px + PX = R_i \end{cases}$$

et, parmi les deux optima relatifs ainsi obtenus, choisit celui qui lui procure la plus grande utilité. La résolution est aisée (cf démonstration dans [7]) et conduit au résultat suivant :

. Si les tarifs rapportés au revenu $\frac{p}{R_i}$, $\frac{A}{R_i}$ sont, dans le plan $(\frac{p}{R_i}, \frac{A}{R_i})$, situés au-dessus de l'arc d'équations paramétriques

$$(4)\begin{cases} \dfrac{p}{R_i} = -\dfrac{v_i'(\xi)}{v_i(0)} & 0 \leqslant \xi \leqslant \infty \\ \dfrac{A}{R_i} = 1 - \dfrac{1}{v_i(0)}\left[v_i(\xi) - \xi v_i'(\xi)\right] \end{cases}$$

où l'on a posé

$$(5) \quad v_i(\xi) = \frac{1}{u_i(\xi, \tau, T_i)} ,$$

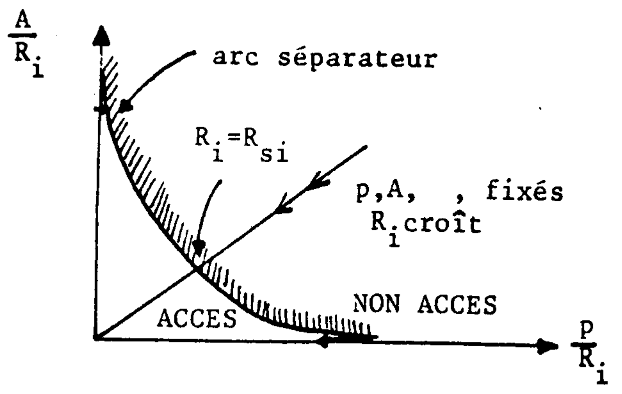
alors le consommateur décide de ne pas accéder (T_i=0) ou de résilier ($T_i > 0$)

. Si le couple $(\frac{p}{R_i}, \frac{A}{R_i})$ se place en dessous de l'arc (4) alors le consommateur accède (T_i=0) ou maintient son accès ($T_i > 0$), et le niveau de consommation est l'unique solution de l'équation implicite en x :

$$(6) \quad \frac{px}{R_i - A} = \frac{-xv_i'(x)}{v_i(x) - xv_i'(x)} ,$$

soit :

$$(7) \quad \overline{x}_i(p, A, R_i, \tau, T_i)$$



L'équation (6) où le bien agrégé "X" est éliminé, traduit la propriété d'isolabilité du bien téléphonique "x" et peut s'interpréter de façon comportementale, comme l'égalisation du budget réellement payé en consommation téléphonique rapporté au revenu net disponible, soit $\frac{px}{R_i - A}$, à un budget relatif désiré, uniquement fonction des préférences de consommation.

On montre aisément à partir de (6) que la consommation $\overline{x}_i$ est fonction décroissante de $\frac{p}{R_i - A}$ c'est à dire croissante de R_i et décroissante de p et A.
L'élimination du paramètre ξ entre les équations (4) de l'arc séparateur des zones d'accès et de non-accès, permet de définir une fonction de revenu seuil :

(8) $R_{si}\ (p, A, \tau, T_i)$

Cette fonction, homogène de degré 1 par rapport aux tarifs p et A représente le revenu minimum que doit posséder l'individu i, caractérisé par l'utilité (1), pour décider de s'abonner (T_i=0) ou de maintenir son abonnement ($T_i > 0$), lorsque les tarifs sont p et A, et le taux de pénétration est τ ; On peut montrer que R_{si}, qui est un indicateur inverse de l'affinité téléphonique, croît lorsque p et A augmentent, décroît lorsque τ augmente.
Notons que la notion de revenu seuil, ici endogène à la modélisation, comme dans [4] , a déjà été employée de façon exogène dans [7] ,[8], et[9] pour modéliser la possession de biens durables.

Nous allons maintenant, en procédant par agrégation, développer successivement un modèle de demande de raccordement à partir de l'expression (8) du revenu seuil R_{si}, puis un modèle de consommation à partir de l'expression (7) de la consommation d'équilibre $\overline{x}_i$.

2. LE MODELE DE DEMANDE DE RACCORDEMENT

2.1. Formulation théorique

Dans cette partie, nous nous intéressons uniquement au modèle de raccordement, pour lequel T_i=0, et non au modèle symétrique de résiliation où la variable T_i, non nulle, peut jouer un rôle explicatif (influence de l'ancienneté sur la probabilité de résiliation). Afin de simplifier l'écriture nous omettrons alors dans la suite de rappeler la variable T_i=0 et nous écrirons par exemple (8) sous la forme :

(8') $R_{si}\ (p, A, \tau)$

Pour procéder à une agrégation sur les individus i, on formulera deux hypothèses "naturelles" de distribution :

i) la distribution en i des revenus R_i est lognormale, de moyenne m et d'écart-type s

ii) la distribution en i des utilités u_i conduit à une distribution lognormale des revenus seuils R_{si}, de moyenne μ et d'écart-type σ ; μ et σ, comme R_{si}, sont d'après (8') des fonctions des tarifs p et A et du taux de pénétration τ ; en particulier, μ (p, A, τ) croît en p et A et décroît en τ. On remarque par ailleurs qu'en raison de l'indépendance , intrinsèque à la modélisation microéconomique, entre fonction d'utilité et revenu, les distributions R_i et R_{si} sont indépendantes.

Avec cette propriété et les hypothèses i) et ii), le taux de demande, c'est à dire le taux de pénétration abonnés + instances τ^{A+I}, s'écrit successivement (cf [6]):

$$\tau^{A+I} = \text{Proba } (R_i > R_{si})$$

$$\tau^{A+I} = \text{Proba } (\text{Log}R_i - \text{Log}R_{si} > 0)$$

Or la distribution $(\text{Log}R - \text{Log}R_s)_i$ étant normale de moyenne $m-\mu$ et d'écart-type $\sqrt{s^2 + \sigma^2}$, on en déduit :

$$(9) \quad \tau^{A+I} = N_{0,1} \left(\frac{m - \mu}{\sqrt{s^2 + \sigma^2}} \right)$$

où $N_{0,1}$ désigne la fonction de répartition normale centrée réduite (•).

Nous supposerons constants les paramètres de dispersion s^2 et σ^2 ; d'après (9), la demande cumulée τ^{A+I} varie alors sous l'effet de la dérive des revenus (évolution de m), sous l'effet des tarifs p et A et sous l'effet du taux de pénétration τ , arguments de la fonction μ(p, A, τ).

μ étant fonction décroissante de τ , τ^{A+I} est une fonction croissante de τ qui admet un seuil de saturation S, atteint lorsque $\tau^{A+I} = \tau = S$ (annulation des instances), et défini par l'équation implicite :

$$(10) \quad S = N_{0,1} \left(\frac{m - \mu(p,A,S)}{\sqrt{s^2 + \sigma^2}} \right)$$

Le taux de saturation S est fonction des tarifs p et A ; il représente pour un jeu de tarifs donné la taille d'équilibre du marché téléphonique au sens de [2] : si une offre volontariste forçait par exemple un taux de pénétration $\tau_0 > S$, alors certains ménages i, ayant un revenu R_i inférieur au revenu seuil requis $R_{si}(p,A,\tau)$, décideraient de se déséquiper, ramenant ainsi le taux vers sa valeur d'équilibre S.

....

(•)

$$N_{0,1}(u) = \int^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} dv$$

Dans le présent modèle, contrairement au cas du modèle logistique classique, le seuil de saturation est déterminé de façon endogène ; cependant, comme dans le modèle logistique, on retrouve une tendance lourde d'évolution de la demande selon une courbe en S, engendrée ici par une intégrale normale, au lieu d'une fonction logistique.

En désignant par M le nombre de ménages, la demande D exprimée sur une période s'écrit :

$$D = M\dot{\Upsilon}^{A+I}$$

soit, en calculant la dérivée temporelle $\dot{\Upsilon}^{A+I}$ à partir de (9) :

$$(11)\quad D = \frac{1}{\sqrt{2\pi}} \exp \left[- \frac{(m-\mu(p,A,\Upsilon))^2}{s^2+\sigma^2} \right] x \frac{1}{\sqrt{s^2+\sigma^2}} \left[\dot{m} - \frac{\delta\mu}{\delta p}\dot{p} - \frac{\delta\mu}{\delta A}\dot{A} - \frac{\delta\mu}{\delta\Upsilon}\dot{\Upsilon} \right]$$

Les effets pris en compte se "lisent" naturellement sur cette équation :

. la "gaussienne" en facteur global correspond à la tendance lourde d'évolution en S de la demande

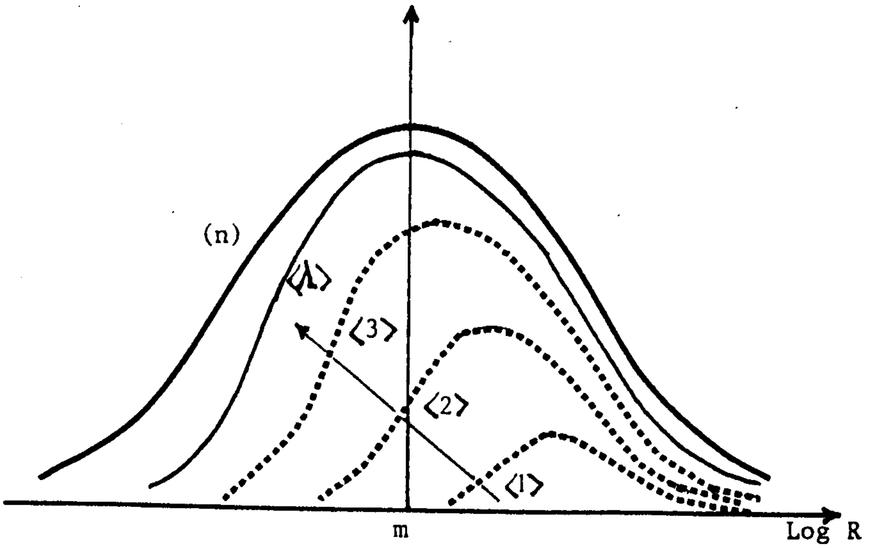
. cette tendance lourde est modulée par une série d'effets additifs :

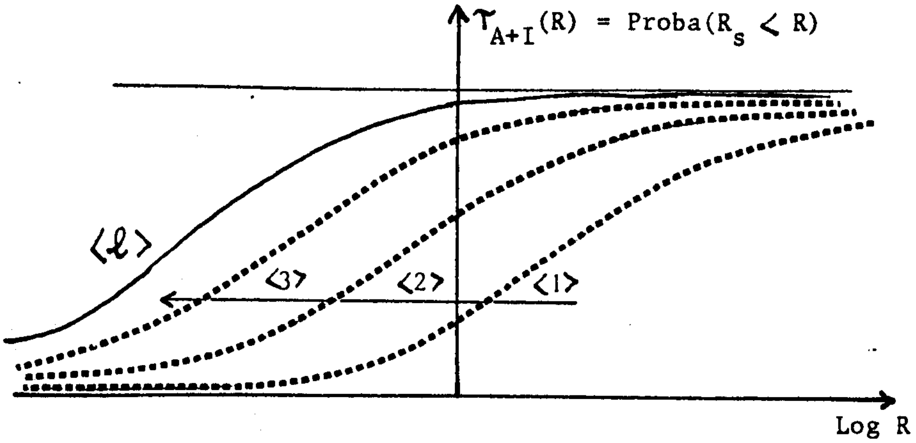
- la croissance de la moyenne des revenus $(\dot{m})$
- l'évolution du prix du trafic $\left(-\frac{\delta\mu}{\delta p}\dot{p}\right)$
- l'évolution des charges fixes $\left(-\frac{\delta\mu}{\delta A}\dot{A}\right)$
- l'externalité de la demande satisfaite $\left(-\frac{\delta\mu}{\delta\Upsilon}\dot{\Upsilon}, \frac{\delta\mu}{\delta\Upsilon} < 0\right)$

L'équation (11) décrit la dynamique de la demande , mais ne renseigne pas sur la composition de cette demande en termes de distribution de revenus ; or on peut montrer en fait que le parc s'enrichit progressivement en abonnés dont la moyenne des revenus est de plus en plus faible. Une illustration visuelle en est fournie par le graphique ci-après où l'on a porté :

. en partie supérieure, d'une part la distribution (normale) des logatithmes des revenus dans l'ensemble de la population (n), et, d'autre part, cette même distribution dans le parc à différentes époques successives $\langle 1 \rangle$, $\langle 2 \rangle$, $\langle 3 \rangle$ (on a négligé la dérive des revenus $\dot{m}$)

. en partie inférieure, également pour les époques successives $\langle 1 \rangle$, $\langle 2 \rangle$, $\langle 3 \rangle$, la répartition de la demande en fonction du revenu $\Upsilon^{A+I}(R)$, qui n'est autre par définition que la fonction de répartition (intégrale normale) du logarithme des revenus seuils (Proba($R_s < R$)) ; cette courbe se décale vers la gauche avec le temps, en raison du glissement de sa moyenne $\mu(p,A,\tau)$, sous l'effet de l'évolution des tarifs et du taux de pénétration ; elle tend à tarifs fixés vers une courbe limite $\langle \ell \rangle$correspondant à la saturation du taux de pénétration (τ = S)





Chaque courbe "supérieure" de distribution des revenus dans le parc$\langle 1 \rangle$, $\langle 2 \rangle$, $\langle 3 \rangle$, s'obtient trivialement en "multipliant" la courbe "inférieure" associée par la courbe normale de distribution des revenus dans la population (n). On observe que la distribution des revenus dans le parc se décale progressivement vers la gauche dans le sens des revenus les plus faibles ; cette distribution tend vers une limite $\langle \lambda \rangle$ associée à $\langle \ell \rangle$et correspondant à la saturation du

Ce résultat sera utile à la modélisation de la consommation (cf § 3.1) car il implique que l'introduction dans le parc de nouveaux abonnés de revenu moyen plus faible et par conséquent d'après (7), consommant moins, tend à faire baisser la consommation moyenne par ligne.

2.2. Spécification économétrique

Les paramètres m et s étant fournis par des enquêtes classiques sur le revenu des ménages, la spécification économétrique de l'équation centrale (11) repose sur une modélisation de μ en fonction de p, A, τ et sur une estimation de σ. Or μ et σ sont accessibles à l'observation, en tant que moyenne et écart type de la répartition de la demande en fonction du revenu, soit $\tau^{A+I}(R)$ (car $\tau^{A+I}(R) = \text{Proba}(R_s < R)$). La distribution $\tau^{A+I}(R)$ est connue sur la période 1971-1978 à travers des enquêtes "conjoncture" et "conditions de vie des ménages" réalisées par l'INSEE (•) avec la collaboration de la Direction Générale des Télécommunications. Un ajustement significatif confirme alors l'hypothèse de lognormalité de la distribution des revenus seuils à condition de segmenter les ménages par catégories socio-professionnelles (CSP), conforte d'autre part la constance de l'écart type σ, et fournit pour μ (moyenne du logarithme des revenus seuils) le modèle suivant, différencié selon chaque CSP n :

$$(12) \quad \mu_n = a_n + b_n \sqrt{\tau_n}\ \text{Log}\tau_n + \underset{(3,0)}{0{,}17\ \text{LogTX}} + \underset{(2,6)}{0{,}76\ \text{LogTB}}$$

où :

.τ_n est le taux de pénétration dans la CSP n
.TX est la taxe de raccordement (en francs constants)
.TB est la taxe de base (en francs constants)
.les coefficients a_n et b_n sont donnés par le tableau ci-après :

(•) Institut National de la Statistique et des Etudes Economiques

C S P n	Constante a_n (T de Student)	Coefficient de $\sqrt{T_n}$ LogT_n b_n (T de Student)
Agriculteurs et salariés agricoles	2,50 (22,1)	-0,10 (17,7)
Patrons industrie et commerce	2,55 (24,8)	-0,09 (11,1)
Professions libérales et Cadres supérieurs	2,82 (20,6)	-0,08 (7,0)
Cadres moyens	2,52 (23,8)	-0,07 (11,9)
Employés et autres actifs	2,64 (23,8)	-0,09 (12,0)
Ouvriers et Personnel de service	2,63 (21,8)	-0,10 (13,4)
Inactifs	2,09 (19,6)	-0,05 (5,1)

Ecart-type de la régression : 2,8 R^2=0,9952 Nombre d'observations:56

Les variables tarifaires explicatives effectivement introduites dans le modèle économétrique, TX et TB, sont liées aux variables théoriques du modèle mathématique, A et p, par les relations :

$$(13)\begin{cases} A = aTX + K.TB \\ p = TB \end{cases}$$

où a est le taux d'actualisation et K le nombre de taxes de base dans la taxe d'abonnement ; en effet, le prix marginal du trafic p est égal à la taxe de base TB, et la charge fixe ressentie A vaut, au moment de la décision d'accès, la taxe d'abonnement K.TB majorée de l'amortissement aTX de la taxe de raccordement.

En désignant par e_t l'élasticité du revenu seuil à la variable tarifaire muette t, les formules de transformation (13) et les résultats de l'ajustement (12) permettent d'écrire

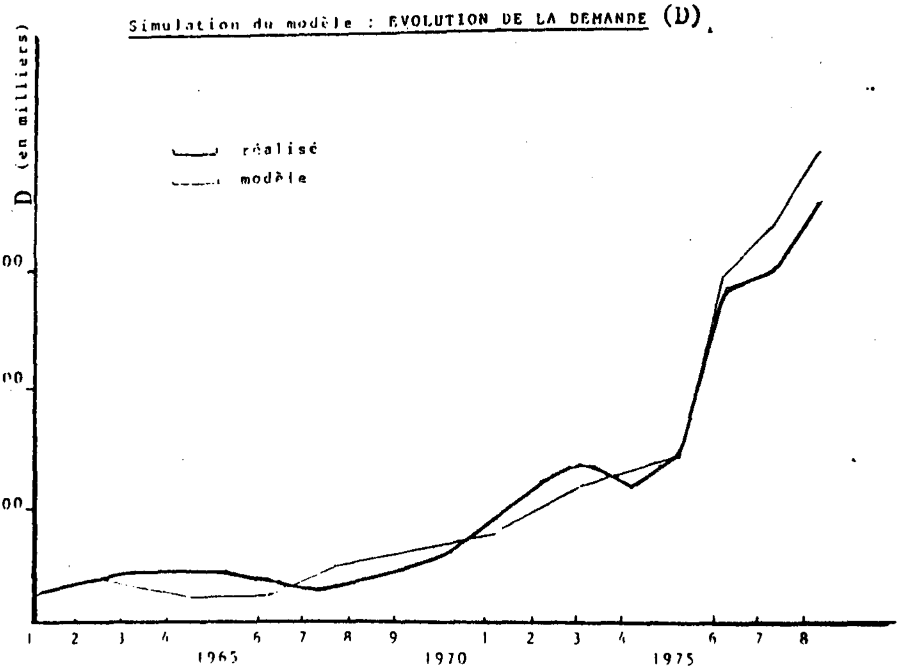
$$(14) \begin{cases} e_{TX} = \dfrac{aTX}{A}\, e_A = 0,17 \\[2ex] e_{TB} = e_p + \dfrac{K.TB}{A}\, e_A = 0,76 \end{cases}$$

La théorie indiquant que le revenu seuil est fonction homogène de degré 1 des tarifs p et A (conséquence des équations (4)), cette propriété doit se traduire par la complémentarité à 1 des élasticités e_p et e_A ; la vérification est convenable puisque de (14) on déduit :

$$(15) \quad e_p + e_A = e_{TX} + e_{TB} = 0,93 \simeq 1$$

D'autre part la condition de positivité de e_p dans le système (14) conduit à établir numériquement (avec TX = 500 F, KTB = 600F/an) que le taux d'actualisation annuel des ménages a ne doit pas être inférieur à 30%, ce qui traduit un comportement de consommation vraisemblable en matière de biens durables.

Le modèle que nous venons de présenter est employé pour la prévision à moyen et long terme (cf [10]). Mis en oeuvre en simulation sur la période 1960-1978 il donne des résultats satisfaisants et explique convenablement en particulier l'"explosion" de la demande de lignes principales nouvelles entre 1974 et 1976 (cf. figure)



3. LE MODELE DE CONSOMMATION

3.1. Formulation théorique

Nous nous intéressons dans ce paragraphe à modéliser l'évolution de la consommation moyenne par ligne résidentielle. Cette consommation moyenne à la date t, soit γ (t), s'obtient par agrégation sur les N(t) lignes du parc P(t) de cette date, des consommations individuelles $\overline{x}_i(p,A,R_i,\tau,T_i)$ issues du modèle microéconomique (cf § 1), selon l'équation :

$$(16)\quad \gamma(t) = \frac{1}{N(t)} \sum_{i \in P(t)} \overline{x}_i(p,A,R_i,\tau,T_i)$$

Cette équation peut être commodément réécrite en regroupant, dans la sommation en i, les individus selon leur date de raccordement $\theta \leqslant t$; tous les individus d'une même "cohorte" "θ" ayant même ancienneté $T_i = t - \theta$, on a :

$$(17)\quad \gamma(t) = \frac{1}{N(t)} \int_{-\infty}^{t} \left[\sum_{i \in "\theta"} \overline{x}_i(p,A,R_i,\tau,t-\theta) \right] d\theta$$

Si N(t,θ) désigne alors l'effectif des lignes de la cohorte "θ" non résiliées à la date t et γ(t,θ) la consommation moyenne par ligne à la date t au sein de cette même cohorte, on a par définition :

$$(18)\quad N(t) = \int_{-\infty}^{t} N(t,\theta)\, d\theta$$

$$(19)\quad \gamma(t,\theta) = \frac{1}{N(t,\theta)} \sum_{i \in "\theta"} \overline{x}_i(p,A,R_i,\tau,t-\theta)$$

et l'équation (17) devient :

$$(20)\quad \gamma(t) = \frac{1}{N(t)} \int_{-\infty}^{t} N(t,\theta)\,\gamma(t,\theta)\, d\theta$$

Par un calcul différentiel simple à partir de (18) et (20), le taux d'évolution de la consommation par ligne, soit $\frac{\dot{\gamma}(t)}{\gamma(t)}$ peut être exprimé sous la forme additive :

$$(21)\quad \frac{\dot{\gamma}(t)}{\gamma(t)} = NA(t) + R(t) + E(t)$$

avec

$$(22)\quad NA(t) = -\frac{N(t,t)}{N(t)}\left[1 - \frac{\gamma(t,t)}{\gamma(t)}\right]$$

$$(23)\quad R(t) = \int_{-\infty}^{t} \frac{-\frac{\delta N}{\delta t}(t,\theta)}{N(t,\theta)} \left[1 - \frac{\gamma(t,\theta)}{\gamma(t)}\right] \frac{N(t,\theta)}{N(t)}\, d\theta$$

$$(24)\quad E(t) = \int_{-\infty}^{t} \frac{\frac{\delta\gamma}{\delta t}(t,\theta)}{\gamma(t,\theta)} \frac{\gamma(t,\theta)}{\gamma(t)} \frac{N(t,\theta)}{N(t)}\, d\theta$$

- NA(t) représente l'effet d'introduction de nouveaux abonnés dans le parc : en effet, ce terme est égal au taux d'apparition d'abonnements nouveaux à la date t, $\frac{N(t,t)}{N(t)}$, pondéré par l'écart relatif, $\frac{\gamma(t,t)}{\gamma(t)}-1$, de la consommation des nouveaux abonnés, $\gamma(t,t)$, à la consommation moyenne des anciens $\gamma(t)$. Or, nous avons vu au paragraphe 2.1 que les nouveaux abonnés ayant des revenus R_i en moyenne inférieurs à ceux des anciens, adoptent des consommations $\overline{x}_i$ en moyenne plus faibles, d'où $\gamma(t,t) < \gamma(t)$, $NA(t) < 0$: l'effet "nouveaux abonnés" est négatif. Plus précisément, cet effet négatif s'explique non seulement par la progression des faibles revenus dans le parc mais aussi par le fait qu'à l'intérieur d'une même tranche de revenu, les nouveaux abonnés consomment moins que les anciens : on peut montrer en effet, sous certaines hypothèses de régularité de la fonction d'utilité téléphonique u_i, que si deux individus ont même revenu $R_i = R_j$ et si leurs revenus-seuil sont tels que $R_{si} < R_{sj}$ - c'est à dire si i s'abonne en premier - alors $\overline{x}_i > \overline{x}_j$, c'est à dire i consomme le plus.

- R(t) représente l'effet des résiliations ; la cohorte des anciens abonnés raccordés au voisinage dela date θ, constituant à la date t la proportion du parc $\frac{N(t,\theta)}{N(t)}\, d\theta$, contribue à ce terme par son taux de résiliation à la date t, $\dfrac{-\frac{\delta N}{\delta t}(t,\theta)}{N(t,\theta)}$, pondéré par l'écart relatif à la moyenne de son niveau de consommation $1 - \dfrac{\gamma(t,\theta)}{\gamma(t)}$; les lignes résiliantes les plus anciennes, consommant plus que la moyenne, ont ainsi un effet négatif sur le taux de variation de la consommation moyenne par ligne et vice-versa.

- E(t) représente l'effet de progression des consommations : en effet, chaque cohorte "θ" contribue à ce terme par le taux d'évolution de sa consommation par ligne $\dfrac{\frac{\delta \gamma}{\delta t}(t,\theta)}{\gamma(t,\theta)}$, pondéré par son "indice" de consommation $\dfrac{\gamma(t,\theta)}{\gamma(t)}$.

Une cohorte consommant comme la moyenne apporte ainsi au taux de variation de la consommation moyenne par ligne $\dfrac{\dot{\gamma}(t)}{\gamma(t)}$ son propre taux brut de variation de consommation par ligne ; cet effet est amplifié (resp. réduit) pour une cohorte ancienne (resp. récente) dont le niveau de consommation $\gamma(t,\theta)$ est supérieur (resp. inférieur) à la moyenne $\gamma(t)$.

Afin de simplifier le modèle, nous ferons l'hypothèse que la probabilité de résiliation d'une ligne i dépend peu de son ancienneté T_i si bien qu'au voisinage d'une date t donnée, toutes les cohortes "θ" ont même taux de résiliation $\mu(t)$ (qui peut être sensible à la conjoncture tarifaire ou économique de la date t), soit :

$$(25) \quad \frac{-\frac{\delta N}{\delta t}(t,\theta)}{N(t,\theta)} = \mu(t) \ , \quad \forall \theta$$

Introduisant (25) dans (23) et utilisant (20), on en déduit

$$(26) \quad R(t) = 0$$

Les résiliations ont donc, sous cette hypothèse, un effet résultant nul sur l'évolution de la consommation moyenne, les résiliations des cohortes récentes de faible consommation compensant exactement celles des cohortes anciennes de plus forte consommation.

Sous cette même hypothèse, on peut écrire d'autre part, en dérivant (18) :

(27) $N(t,t) = \dot{N}(t) + \mu(t)\, N(t)$,

d'où l'effet "nouveaux abonnés" :

(28) $NA(t) = -\left[\frac{N(t,t)}{N(t)} + \mu(t)\right]\left[\ 1 - i(t)\ \right]$,

où l'on a noté :

(29) $i(t) = \frac{\gamma(t,t)}{\gamma(t)}$,

l'indice de consommation d'un nouvel abonné.

Aux résiliations près, $-(1 - i(t))$ apparaît ainsi dans (28) comme l'élasticité de la consommation $\gamma(t)$ à l'effectif du parc $N(t)$. De plus on déduit de façon théorique du modèle de demande que l'indice $i(t)$ (resp. l'élasticité $|1-i(t)|$) est décroissant (resp. croissante) et tend vers une limite. En effet la distribution des revenus dans le parc se décalant vers les revenus inférieurs lorsque le taux de pénétration augmente (cf § 2.1), la moyenne des revenus donc la moyenne des consommations des abonnés les plus récents, sont de plus en plus faibles relativement à ces mêmes moyennes calculées sur l'ensemble des abonnés, l'écart se creusant jusqu'à une limite positive associée à la saturation S du parc. Nous verrons au § 3.2 que cette propriété de stabilisation de l'indice $i(t)$ admet une bonne vérification empirique.

Enfin l'écriture du terme d'évolution $E(t)$ peut être simplifiée en partageant, d'après (19), le trend de consommation de chaque cohorte "θ", soit $\frac{\delta\gamma}{\delta t}(t,\theta) \,/\, \gamma(t,\theta)$, en deux composantes, l'une, $g(p,A,\tau)$, associée aux variables de tarifs p, A et de taux de pénétration τ, l'autre, $f(t-\theta)$, associée à la variable d'ancienneté $T_i = t - \theta$; d'où l'expression :

(30) $$\frac{\frac{\delta\gamma}{\delta t}(t,\theta)}{\gamma(t,\theta)} = g(p,A,\tau) + f(t-\theta),$$

puis en remplaçant (30) dans (24) et en utilisant (20), il vient :

$$(31)\quad E(t) = g(p,\, A, \tau) + \int_{-\infty}^{t} f(t-\theta)\ \frac{\gamma(t,\theta)}{\gamma(t)}\ \frac{N(t,\theta)}{N(t)}\ d\theta$$

On mettra effectivement en évidence au § 3.2 que les cohortes les plus récentes ($t-\theta \leqslant 6$ ans) ont,en excès sur le trend de base $g(p,A,\tau)$,un coefficient de trend supplémentaire $f(t-\theta)$ positif, conduisant à un rattrapage partiel de la consommation des nouveaux abonnés sur celle des anciens (cf [11] et [12]). Le coefficient $f(t-\theta)$ considéré au niveau de chaque abonné, indique une montée en charge du trend de consommation sur une période de 6 ans environ à partir de la date d'abonnement ; ce phénomène de montée en charge renvoie dans le modèle microéconomique du §1 à l'évolution de la fonction d'utilité u_i sous l'influence de la variable d'ancienneté T_i ; cette évolution doit alors s'interpréter,non pas comme un processus d'apprentissage de "l'outil téléphone", qui serait un effet à court terme,mais plutôt comme une extension progressive de la part affectée au téléphone dans la réalisation des tâches de communication du ménage abonné.

D'après (21),(26), (28) et (31), l'évolution de la consommation moyenne par ligne est finalement régie par l'équation différentielle suivante :

$$(32)\quad \boxed{\frac{\dot{\gamma}(t)}{\gamma(t)} = -\left[\frac{\dot{N}(t)}{N(t)} + \mu(t)\right]\left[1-i(t)\right] + g(p,A,\tau) + \int_{-\infty}^{t} f(t-\theta)\,\frac{\gamma(t,\theta)}{\gamma(t)}\,\frac{N(t,\theta)}{N(t)}\,d\theta}$$

où les termes $\gamma(t,\theta)$, $N(t,\theta)$, figurant dans le terme intégral sont eux-mêmes engendrés par les équations dynamiques (25) et (30).

L'équation centrale (32) résume clairement les trois effets additifs contribuant à faire évoluer la consommation par ligne résidentielle :

- effet de démographie du parc, ou effet nouveaux abonnés NA(t)
- trend de consommation de base commun à l'ensemble du parc, $g(p,A,\tau)$, variable avec les tarifs et le niveau de pénétration
- différenciation des trends de consommation selon l'ancienneté d'abonnement du fait de la montée en charge (terme intégral en $f(t-\theta)$)

L'économétrie montrera en fait (cf § 3.2) que la dynamique de consommation résulte essentiellement des deux premiers effets (le premier jouant négativement le second positivement), tandis que le troisième effet,plus faible,intervient comme une correction.

3.2. Spécification économétrique

La spécification économétrique de l'équation (32) repose alors sur :

- l'ajustement du taux de résiliation $\mu(t)$
- l'ajustement de l'indice de consommation des nouveaux abonnés i(t)
- la modélisation de la fonction de trend $g(p,A,\tau)$
- l'ajustement du coefficient d'ancienneté $f(t-\theta)$

Tous les ajustements sont réalisés par région ; les méthodes employées et les résultats obtenus sont les suivants :

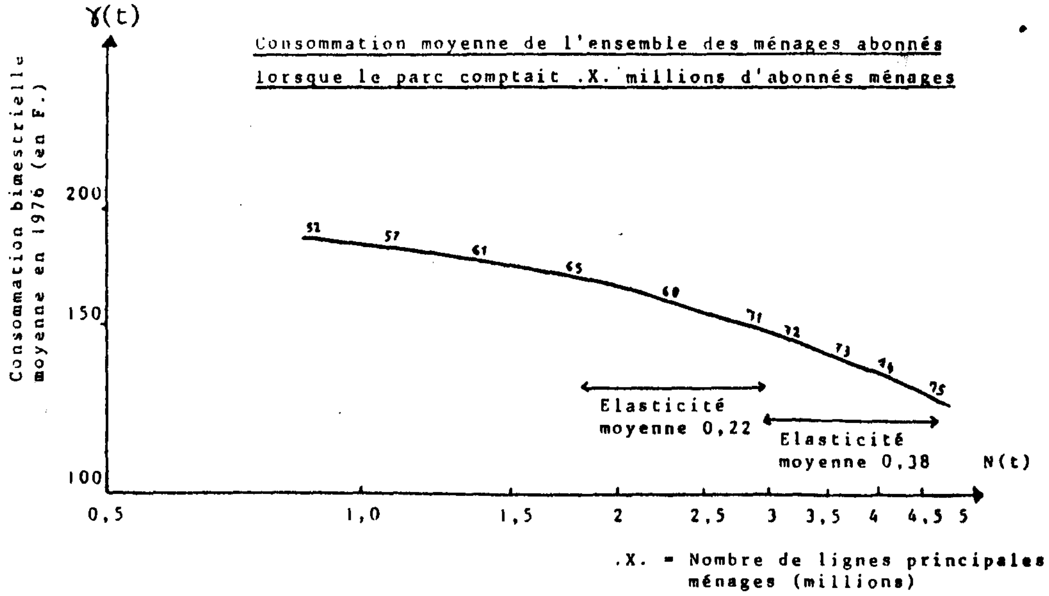
a. taux de résiliation $\mu(t)$

En moyennant les valeurs relativement stables observées dans le passé, on a retenu une valeur constante par région selon le tableau suivant :

REGION	μ (%)
NORD	1,2
EST	1,3
CENTRE EST	1,1
SUD EST	1,4
SUD	1,3
SUD OUEST	1,3
OUEST	1,2
CENTRE OUEST	1,2
ILE DE FRANCE	3,0

b. indice de consommation des nouveaux abonnés i(t)

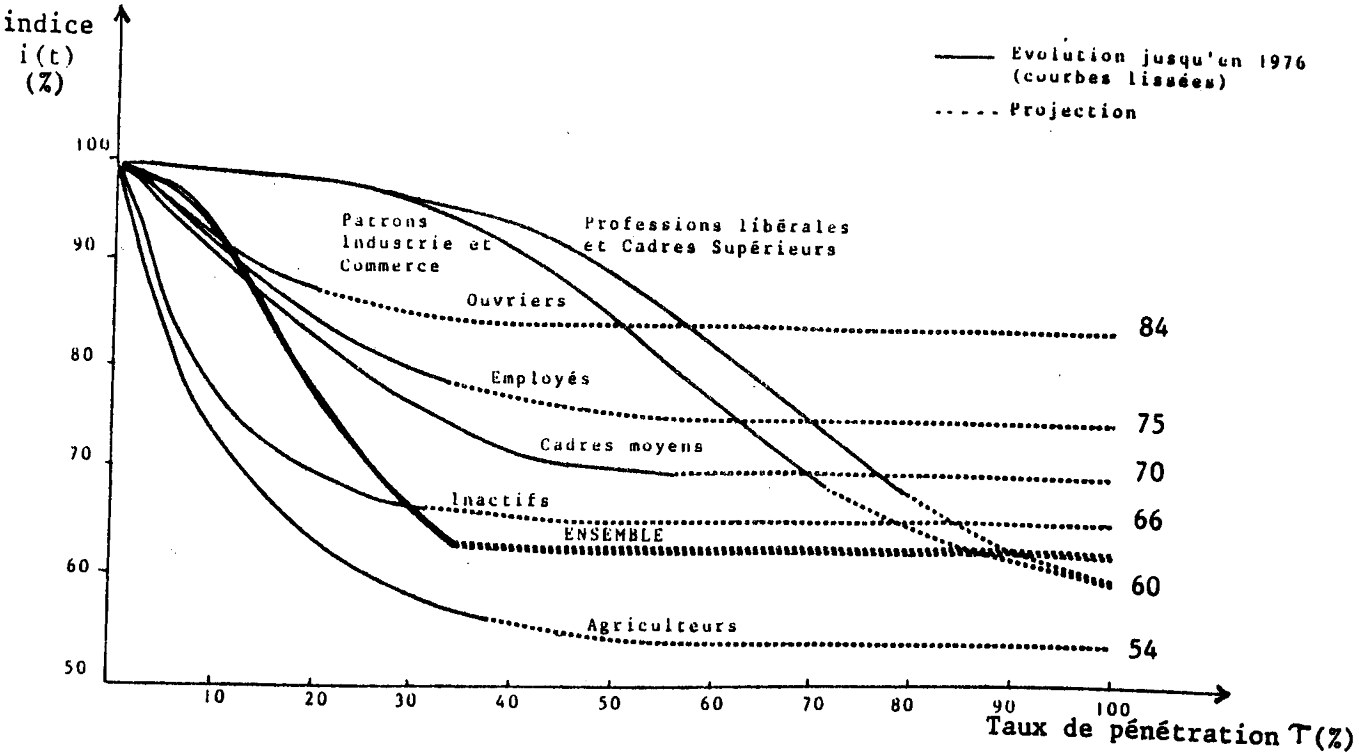
L'historique de cet indice au niveau national a d'abord été reconstitué en figurant l'évolution fictive qu'aurait connu la consommation par ligne si les différentes "cohortes" d'anciens abonnés avaient gardé un niveau de consommation constant (celui de 1976). L'allure de la courbe ainsi obtenue indique comment varie la consommation par ligne sous l'effet de la seule progression du parc, l'effet de trend étant éliminé ; sa pente représente exactement l'élasticité de la consommation par ligne à l'augmentation du parc. On observe que cette élasticité a fortement varié dans le passé : ainsi, elle vaut en moyenne -0,22 sur la décennie 61-71 puis en moyenne -0,35 sur la période 71-75.



L'indice de consommation des nouveaux abonnés qui, d'après l'équation fondamentale d'évolution de la consommation (32), est égal à l'élasticité au parc augmentée de 1, a donc diminué progressivement dans le passé vers sa valeur actuelle, voisine de 0,7.

Cependant ,malgré cette décroissance passée, une hypothèse de constance de l'indice peut être retenue pour la prévision : l'observation, en accord avec la prédiction théorique (cf § 3.1), montre en effet qu'à l'intérieur de chaque CSP (Catégorie Socio-Professionnelle) l'indice se stabilise nettement vers une valeur asymptotique. On peut alors simuler son évolution future pour l'ensemble des ménages en pondérant doublement les indices de chaque CSP par la part variable de cette CSP dans la dérive du parc et par son niveau de consommation rapporté au niveau moyen des ménages. On constate ainsi que l'indice global s'écarte peu de la valeur 0,7 que nous maintiendrons donc constante dans le modèle.

Evolution et Projection de l'indice de consommation des nouveaux abonnés par CSP



En fait l'évaluation de l'indice est faite au niveau régional et on observe une certaine variance géographique selon le tableau ci-dessous :

REGION	INDICE DE CONSOMMATION DES NOUVEAUX ABONNES RESIDENTIELS
NORD	0,70
EST	0,65
CENTRE EST	0,72
SUD EST	0,74
SUD	0,75
SUD OUEST	0,76
OUEST	0,70
CENTRE OUEST	0,67
ILE DE FRANCE	0,80

c. <u>fonction de trend de consommation</u>

La méthode d'ajustement consiste à isoler dans chaque région n un effectif constant d'abonnés extraits d'un échantillon vivant de lignes téléphoniques - le PANEL (cf [13][14])- et à observer l'évolution de la consommation de ce sous-échantillon de taille fixe ; dans la pratique, on a "suivi" bimestre par bimestre la consommation par ligne $\gamma_n(t)$ des abonnés raccordés avant 1974 et on a pu spécifier une relation du type :

$$(33)\ \mathrm{Log}\ \gamma_n(t) = a_n t + b\ \mathrm{Log\ TB} + c\ \mathrm{Log\ IPI} + \mathrm{cste}$$

où t est le temps, TB la taxe de base, IPI l'indice de la production industrielle.

Par rapport au modèle mathématique théorique où le trend de consommation est engendré par les variables p, A et τ, à travers la fonction $g(p,A,\tau)$, le modèle empirique (33) présente quelques différences :

- le taux de pénétration τ y est remplacé par le temps t, fortement correlé avec τ
- la taxe de base TB résume à la fois le prix du trafic p=TB et la taxe d'abonnement A, ce qui correspond au fait que p et A variaient parallèlement et proportionnellement à TB sur la période d'ajustement (A = K.TB)
- on a introduit l'indice IPI afin de tenir compte des effets de conjoncture.

Le trend de base a_n est d'environ 10%,avec les modulations régionales suivantes:

REGION n	TREND DE CONSOMMATION a_n
NORD	10,0
EST	9,4
CENTRE EST	8,0
SUD EST	7,5
SUD	10,6
SUD OUEST	10,3
OUEST	8,8
CENTRE OUEST	9,5
ILE DE FRANCE	10,0

Ces valeurs seraient assez fragiles si, a posteriori, la bonne adéquation du modèle de prévision avec la réalité sur la période 68-77 ne plaidait pour leur stabilité relative dans le temps. Cependant, pour la prévision à l'horizon 1985 il apparaît raisonnable de postuler un ralentissement de la croissance de la consommation à parc constant et donc un léger fléchissement des trends, correspondant à un alignement sur les trends observés dans les autres pays européens. Une étude en cours vise à rendre compte d'une telle évolution des trends en introduisant explicitement dans le modèle (33) le taux de pénétration τ(rapprochant ainsi le modèle économétrique du modèle mathématique) ; en effet, la saturation de τ, en créant du même coup une saturation de l'externalité associée, devrait provoquer à terme un ralentissement de la croissance de consommation. Une telle approche est compatible avec les observations actuelles,la croissance régulière constatée de la consommation pouvant s'expliquer par une croissance du parc elle-même encore régulière au voisinage de l'inflexion logistique.

L'élasticité b à la taxe de base n'a pas été trouvée significativement différente de zéro, mais ce résultat est peu fiable en raison de la forte variabilité des séries de consommation et du faible nombre de changements tarifaires intervenus au cours de la période d'estimation. En outre, la nouvelle approche de modélisation en cours (introduction explicite du taux de pénétration τ) semble fournir une valeur significativement positive de b. Si tel est le cas, le modèle microéconomique permet alors de prédire que cette valeur b correspond essentiellement à une élasticité au prix du tarif p et non à la charge fixe d'abonnement A (ce qui est actuellement indécidable économétriquement ces deux tarifs ayant évolué proportionnellement sur la période d'ajustement). En effet, les consommations individuelles $\overline{x}_i$ étant uniquement fonction de $(R_i-A)/p$ (cf §1.2),il est aisé de montrer que leurs élasticités à p sont beaucoup plus fortes que leurs élasticités à A, pourvu que l'abonnement A soit négligeable devant le revenu R_i. Dans l'attente d'une confirmation empirique, ce résultat théorique sera utile à la prévision puisque les évolutions de la taxe de base et de la taxe d'abonnement sont dissociées depuis Juin 1979.

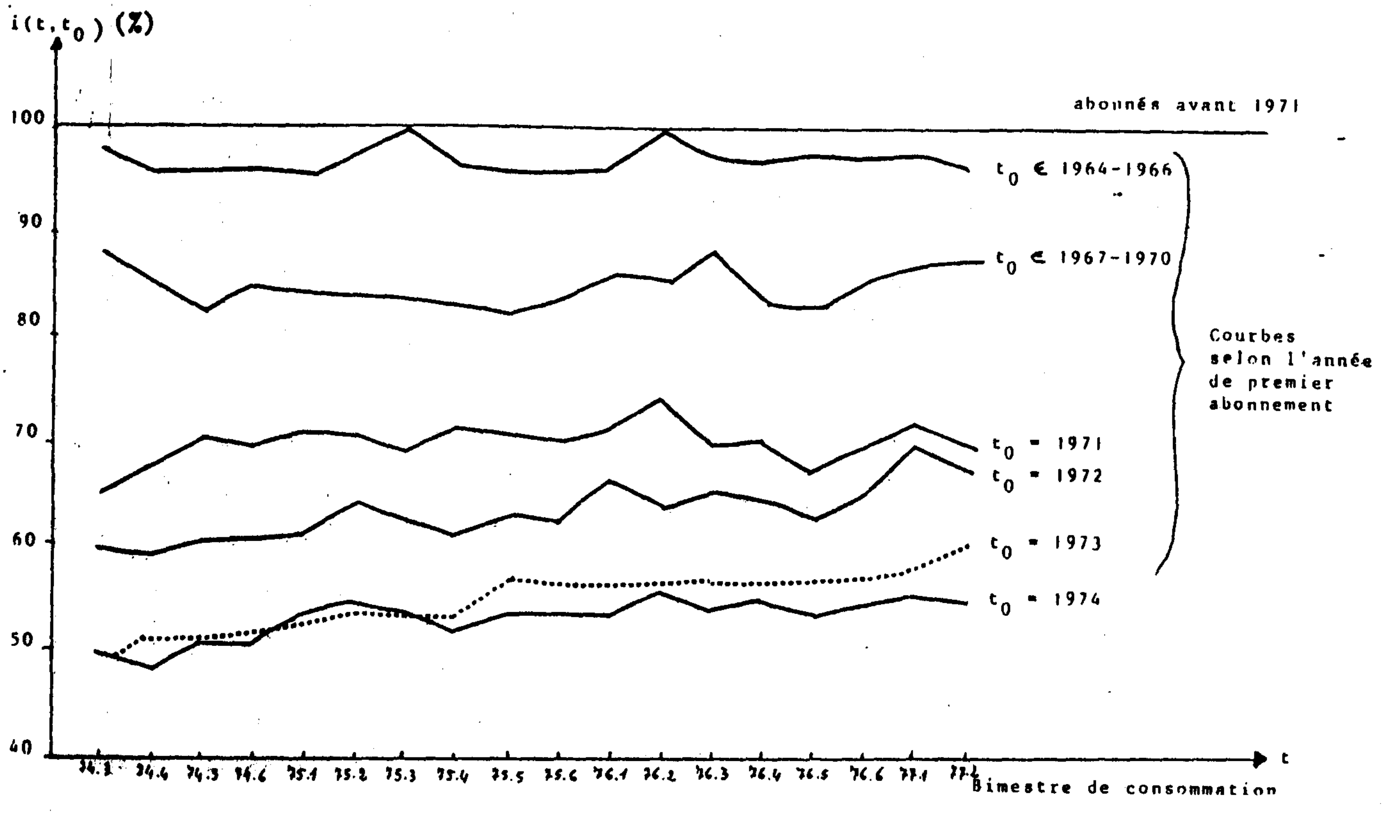
Les données issues du PANEL portent sur une période trop courte pour que l'on puisse mettre directement en évidence l'élasticité à l'IPI. On a donc estimé cette élasticité indirectement, sans travailler à parc constant, à partir de statistiques de consommations globales ; elle est voisine de 0,1.

d. coefficients d'ancienneté $f(t-\theta)$

On cherche à ajuster les coefficients d'ancienneté ou coefficients de montée en charge $f(t-\theta)$, composantes venant s'ajouter au trend de base de la consommation pour les cohortes d'abonnés récemment raccordés. Ces coefficients, faibles en valeur absolue sont difficiles à estimer directement avec précision à partir du trend observé pour chaque cohorte d'abonnés car ce trend est essentiellement constitué par la composante de base a $\simeq$10%, la composante tarifaire (b Log TB) et la composante conjoncturelle (c Log IPI). C'est pourquoi, afin "d'éliminer" ces composantes "lourdes", nous avons cherché plutôt à obtenir $f(t-\theta)$ comme la différence entre les trends de consommation de deux populations d'abonnés distinctes :on a considéré à cet effet, d'une part les abonnés raccordés avant 1971, et d'autre part ceux qui ont été raccordés à une date postérieure, t_0, donnée entre 1971 et 1974.

Soit maintenant $i(t,t_0)$ l'indice qui, à chaque date t postérieure à 1974, rapporte la consommation de la seconde population à celle de la première. Le taux d'évolution de cet indice en fonction de la date courante t est une mesure du taux de rattrapage des abonnés raccordés en t_0 sur ceux présents dans le parc avant 1971 ; ce taux de rattrapage s'identifie exactement au coefficient de montée en charge $f(t-t_0)$ sous l'hypothèse que la durée de montée en charge n'excède pas 4 ans. (Si tel n'était pas le cas en effet, les abonnés de 1971 seraient encore en montée en charge au cours de 1974 et le taux de rattrapage serait égal à la montée en charge des abonnés de t_0, soit $f(t-t_0)$ diminuée d'un terme correctif complexe exprimant la montée en charge d'abonnés antérieurs à 71).
Le graphique ci-dessous représente plusieurs valeurs de t_0, l'évolution de l'indice $i(t,t_0)$ en fonction de t.

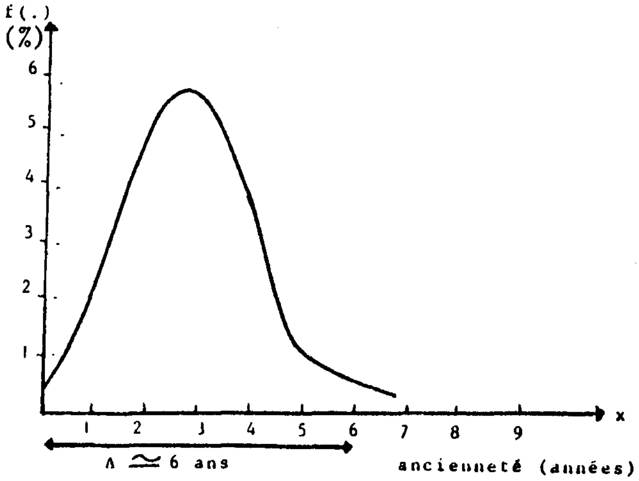
Evolution de l'indice de consommation $i(t,t_0)$ des ménages suivant l'année de premier abonnement t_0 (100 = abonnés avant 1971



On remarque que :

. d'une part, il existe un certain rattrapage à long terme des abonnés raccordés à une date donnée (1971) sur les "anciens" raccordés avant cette même date ; en effet, les abonnés raccordés en 1971 avec un indice de consommation initial de 0,6 atteignent l'indice 0,7 en 1977

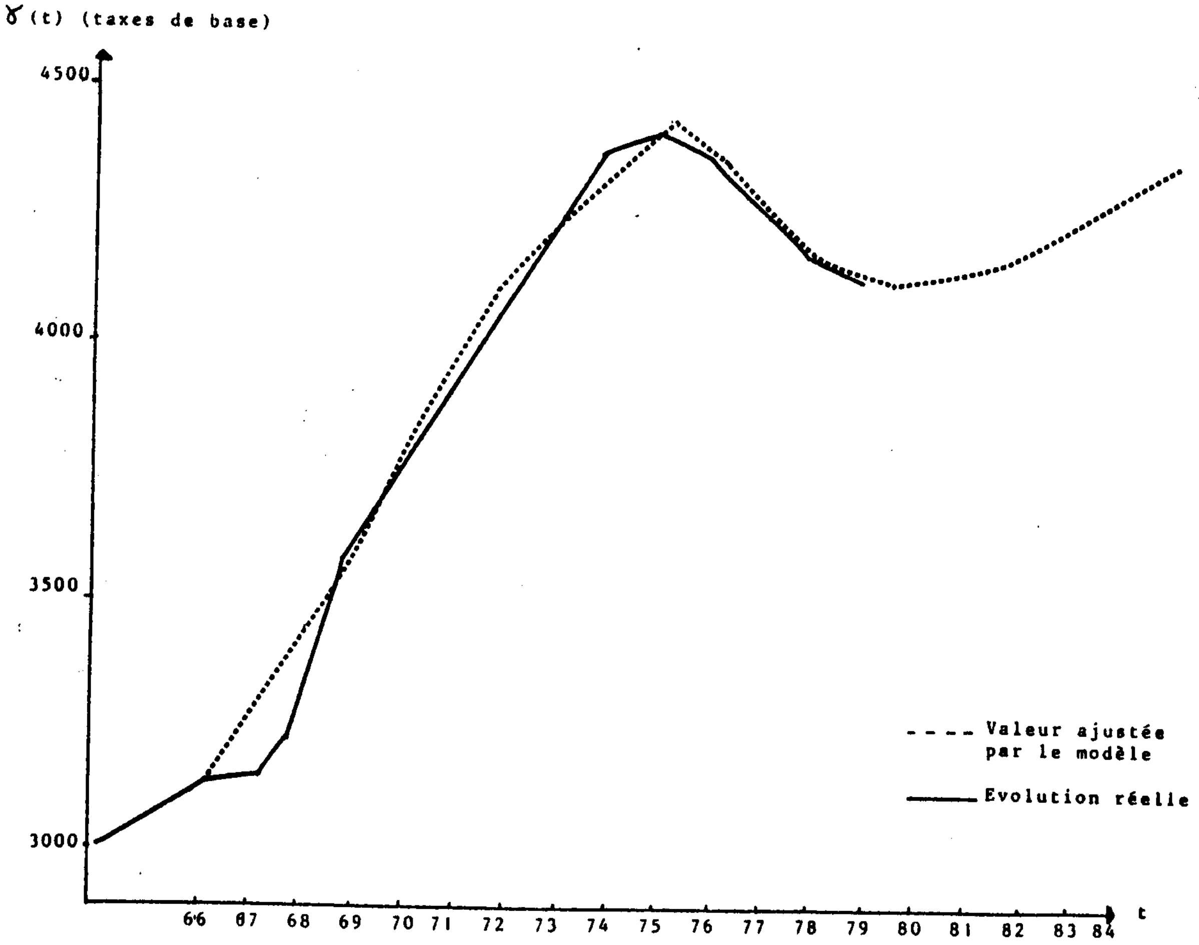
. d'autre part, pour t_0 postérieur à 1971, l'indice $i(t,t_0)$ croît le plus vite en fonction de t pour les abonnés les plus récents (t_0 = 73 ou 74). Sur chaque courbe du graphique, caractérisée par une valeur de t_0, on a appliqué une régression exponentielle et on a porté sur la figure ci-dessous, en ordonnée, les taux de croissance annuels de l'indice ainsi obtenus en moyenne sur la période mi 74 - début 77, et en abscisse, les anciennetés moyennes correspondantes $75-t_0$. On a ainsi approximativement représenté la fonction de montée en charge f(.).



Il semble donc qu'il existe une montée en charge lente de la consommation des abonnés ménages avec l'ancienneté s'étalant globalement sur une période de 5 à 7 ans et présentant un maximum entre 2 et 3 ans.

L'hypothèse préalable à l'estimation, selon laquelle la durée de montée en charge ne devait pas excéder 4 ans, n'étant pas rigoureusement vérifiée, un calcul simple de majoration montre alors que l'erreur relative commise de ce fait dans l'évaluation des coefficients est en tout cas inférieure à 10% (cf [12]).

Le modèle de consommation résidentielle que nous venons de présenter, complété par un modèle professionnel (cf [12]) rend très bien compte de l'évolution passée par rétropolation jusqu'en 1966, en-deça de sa période d'ajustement (74-79) (cf figure)



Il prévoit entre 1980 et 1985 une reprise de la croissance de la consommation par ligne au taux moyen de 1,2% essentiellement due :

. d'une part au maintien de l'effet soutenu des trends, même dans l'hypothèse d'un léger fléchissement de ces trends correspondant à un alignement sur la moyenne des trends observés dans les autres pays européens

. d'autre part à une stabilisation de la croissance de la part relative des ménages dans la dérive du parc au voisinage de la saturation ; en effet la forte croissance de cette part relative sur la période 75-80 a produit un effet négatif très important sur la consommation moyenne, la consommation par ligne des ménages étant très inférieure à celle des entreprises.

BIBLIOGRAPHIE

[1] ARTLE R. and AVEROUS C.

(1973) "The telephone System as a public good : Static and dynamic aspects"
The Bell Journal of Economics and Management Science, Vol 4.

[2] ROHLFS J.

(1974) "A theory of interdependent demand for a communications service"
The Bell Journal of Economics and Management Science, Vol 5.

[3] SQUIRE L.

(1973) "Some aspects of optimal pricing for telecommunications"
The Bell Journal of Economics and Management Science, Vol 4.

[4] TAYLOR L.

"Telecommunications demand : a survey and critique"
Ballinger Company Cambridge, Massachussets 1980

[5] GAMOT G.

"Théorie microéconomique"
Cours professé à l'Ecole Nationale Supérieure des Télécommunications

[6] CURIEN N.

"La consommation télephonique : une approche microéconomique".
3me Congrès international sur l'Analyse, la Prévision et la Planification
dans les Services Publics Juin 1980

[7] CURIEN N.

"Le téléphone et les biens à accès : modèles de demande et de consommation"
Note Direction Générale des Télécommunication Septembre 1980.

[8] RAULT C.

"Etude économétriquede la possession d'un ensemble de biens durables de consommation"
Annales de l'INSEE n° 1

[9] VON RABENEAU and STAHL
(1974) "Dynamic aspect of public goods : a further analysis of the telephone system" ,
The Bell Journal of Economics and Management Science Vol 5, Autum, 1974

[10] GENSOLLEN M. , VILMIN E.

"Prévision de la demande téléphonique"
3me Congrès international sur l'Analyse, la Prévision et la Planification
dans les Services Publics Juin 1980

[11] CURIEN N., DANG NGUYEN G.

"Prévision du produit du trafic téléphonique"
3me Congrès international sur l'Analyse, la Prévision et la Planification
dans les Services Publics Juin 1980

[12] VILMIN E., CURIEN N.
"La consommation téléphonique : prévisions à l'horizon 85"
Note Direction Générale des Télécommunications Juin 1979

[13] BERTHON D., CHABROL J.L. , LEROY F., WERKOFF M.
" Le Panel des abonnés au téléphone "
3me Congrès international sur l'Analyse, la Prévision et la Planification dans les Services Publics Juin 1980.

[14] BERTHON D., CURIEN N.
"Le plan de sondage utilisé pour le Panel des abonnés au téléphone"
Note Direction Générale des Télécommunications Octobre 1980

o o

o

B.C./ALBERTA LONG DISTANCE CALLING

Alain de Fontenay
Department of Communications
Government of Canada

J.T. Marshall Lee
British Columbia Telephone Company

March, 1981 (Revised)

This paper represents the view of the authors,
not necessarily those of the Government of Canada or
the British Columbia Telephone Company

The authors would like to thank Jon Breslaw, Debbie Gorham and John Routledge for helpful comments and discussions on an earlier draft.

1. Introduction and Outline

In the last ten years there have been a large number of telephone demand studies on both theoretical and empirical levels. From these studies two observations are in order: first, analysts appear to be very divided with respect to the significance of network externality, and second, most studies simply invoke the constant price and income elasticities assumption by estimating a double-log equation without first investigating the validity of the assumption. In light of these two observations, the purpose of this study is (1) to critically discuss the presence of network externality and other related issues, (2) to investigate and verify the validity of the commonly employed constant elasticity assumption, and finally (3) to explore the use of pooled time series and mileage band data in obtaining a better and more general demand equation for toll where price and income elasticities may vary. For this purpose, a common data base (B.C. to Alberta Monday to Friday day DDD residence call minutes from 1973 - 1979 Q4), is set up to evaluate alternative specifications and competing hypotheses as well as to estimate the proposed demand equation.

Taylor (1980) has presented a thorough and enlightening review of the empirical demand literature. Starting from Taylor, the second part of this paper reviews the subsequent literature in Canada. Particular attention is given to the estimation of toll elasticities. Issues relating to dynamic adjustment processes, normalization and degree of disaggregation are also considered.

It is also argued in this section that the assumption of constant price elasticity is questionable. In most empirical studies where calls are disaggregated by mileage bands, the common finding is that price elasticities in absolute value increase with distance. Since the price of a call usually increases with distance, the above finding implies that price elasticities actually increase with prices. The translog demand function is later introduced to cope with this observation.

In section 3, the issue of consumption and network externalities is discussed and illustrated using B.C. to Alberta calling data.

In section 4, the translog demand function is introduced to cope with the empirical observation cited earlier that price elasticities in absolute value increase with distance.

The estimated results of the translog demand function are presented in this section 5. Pooled time-series mileage bands data are used in the estimation. The pooling of data is based on the presumption that consumer's welfare is independent of distance, distance affecting solely the price and opportunity cost of the call. Price elasticities for each mileage band are computed and compared with results from the double log specification for reasonableness.

Finally, in part IV, some concluding remarks are offered.

2. Modelling Message Toll Demand: A Constructive Review of the Literature

2.1 The Double Log (DL) Model

In this section the Canadian literature subsequent to Taylor (1980) in the area of the empirical analysis of demand for message toll services is reviewed. This literature is divided between work done by carriers (Bell Canada, 1976; Dreessen, 1977; Dreessen, 1979; Bell Canada, 1980; Piekaar, 1980) and work done at various universities, often under the Department of Communications' sponsorship, (Bernstein et al., 1977; Corbo et al., 1978; Fuss and Waverman, 1978; Corbo et al., 1979; Breslaw and Smith, 1980; Breslaw, 1980; Bernstein, 1980[1]; Breslaw and Smith, 1981 (a) and (b) and Fuss and Waverman, 1981).

Two factors characterize this literature, namely its quasi-universal adoption of double-log (DL) types of demand models and the sharp differences in the treatment of the network externality and in the magnitudes of estimated elasticities. To illustrate these differences, it suffices to note that elasticity estimates are ranging from a low of -.18 in Bell Canada (1980)'s estimate for mileage bands of 100 miles and under to highs in the neighbourhood of -1.3 obtained at Concordia University (see for instance Breslaw and Smith, 1980) or at the University of Toronto (see for instance Fuss and Waverman, 1981).

In spite of the general acceptance of the DL model, analysts differ in the particular specifications they have adopted. These may involve various aggregations, the time structure of the model, the variables used, etc. For instance, industry studies have been by far the most disaggregated, both in terms of the message toll services outputs considered and in terms of time. In fact, the Intra B.C. model and Bell Canada (1976) are based on monthly series. These differences in specification will be reviewed, as will be the concept of network externality, and conclusions will be drawn regarding the desired specification of a demand model.

While the strength of the DL model lies in its very simplicity, its justification has been rather scanty. It has been derived through the Box-Cox transformation in Corbo et al. (1979) (see also Fuss and Waverman, 1981), and Breslaw and Smith (1981, (a) and (b)) have gone to great length to set it within a utility maximization framework. Alternative forms have been explored, hence, Bernstein et al. (1977) considers the Rotterdam model while Corbo et al. (1979), investigates the application of flexible functional forms such as the translog and the generalized Leontief. To date, those efforts have not been successful, and the DL model appears more entrenched than ever.

In this section, a new specification for message toll demand is proposed. The aim is to cope with the empirically observed relationship between the message toll price elasticity and distance which will be hypothesized to be the result of an ex ante dependence between the elasticity and the price. Unlike Breslaw and Smith (1981), the relationship between the utility maximization process and the specification of the demand curve shall not be investigated as such a process would, under most common specifications, imply additional constraints which have not been considered. Instead, we will introduce the direct translog demand function (not to be confused with the demand function which can be derived as a first order condition from a translog utility function; Corbo et al., (1979). Simply, this function is a second order approximation to an arbitrary demand curve in which the variables are expressed in terms of the logarithms. Hence, it is of the form of a translog function, and a straightforward generalization of the DL model.

2.2 The Aggregation Problem

Outside of the industry studies, a common flaw among the Canadian literature appears to be related to the data base used. All these studies, with the exception of part of Bernstein et al. (1977) and the whole of Bernstein (1980), are based on the same data base, Olley's Bell Canada Productivity Study data base (Bell Canada, 1969, 1973 and 1980). This data base disaggregates Bell Canada's output into local, three categories of message toll services, other toll services and miscellaneous revenues. In the revised version of the productivity data base first made available by Bell Canada in 1980, other toll is itself disaggregated in WATS, TWX, and Private Lines. Message toll service is disaggregated into Intra-Bell, TCTS and adjacent, and US and overseas. All of these studies but Breslaw (1980) reaggregate those series through a Torqvist index in a message toll output series, and use this aggregate series to estimate the demand for message toll services. Breslaw (1980) also considers individually the demand for each of WATS, Intra-Bell, TCTS and adjacent, and US and overseas.

Certain observations regarding the applicability of the productivity data base to demand analysis are in order. First, these series are deflated settled revenues; only accidentally would settled revenues correspond to the output measures which correspond to the quantity demanded.[2] The relevant output measure in a study of subscribers' demand is the output which corresponds to calls originated by subscribers (even though the relevant measures would be more complex whenever Taylor's call externalities are introduced) and the revenue measure to which it corresponds is the originating revenue. It should be noted in passing that the price indexes used as deflators are indeed chained price indexes, based on originating revenue.[3,4] The industry studies are free from that flaw since they use deflated originating revenues (Bell Canada), or actual quantities (B.C. Tel).

Another major flaw of most studies is their level of aggregation. As noted earlier, among those studies based on Bell Canada's productivity data base, only Breslaw (1980) considers the disaggregation between classes of message toll services while Fuss and Waverman's (1981) study, in one of their models, considers a disaggregation by mileage bands. Bell Canada's study is restricted to Intra-Bell message toll services, which is disaggregated into two categories depending upon whether or not the distance is greater than 100 miles. B.C. Tel's analysis is by far the most disaggregated since not only the Intra-B.C. Tel category alone is considered, but it is divided further into seven mileage bands.

The aggregation problem is extremely serious in view of the simple observation that in almost all models, estimates of demand elasticity increases with the mileage, an observation already made by Taylor (1980) and Fuss and Waverman (1981).[5]

The aggregation problem appears at two levels: those of the time and output characteristics. Industry studies are based on monthly or quarterly series, over relatively short time periods, typically 6-7 years. All the other analyses but for one are based on yearly series and cover a much longer time span, typically over twenty five years. The rationale for the rather short period used in industry studies is typically based not only on data availability but also on the presumption that the one-minute minimum duration was introduced in the early seventies and that the access market reached saturation around that period, the net effect being a structural change in the beginning of the seventies.

A message toll call can be indexed in terms of its class (intra, adjacent, etc.), its originating customer (business, resident), its type (DDD, SOH, PP), its time of day (TOD), (day, evening, night) - day of week (week day, week-end and holidays) characteristics, its duration and its distance. Only the B.C. Tel's study takes into account the complete disaggregation in terms of types of call, TOD and distance. Even though one would expect that message and duration are determined simultaneously, one would expect that the absence of an access charge, in the form of a higher price for the first minute, should minimize the desirability to go beyond the message - minute as the output unit. The extent of the disaggregation, in the Intra B.C. model, is such as to raise some estimation problems (Piekaar, 1980).

Bell Canada restricted its attention to DDD in terms of two mileage bands but regardless of TOD (also excluding holidays), duration and type of customer, and to person-to-person with the same restrictions, but without mileage bands differentiation. This led Bell Canada to use deflated revenue output measures. Only Fuss and Waverman (1981), in one of their attempts, consider duration explicitly. Finally, Bernstein (1980) uses messages as his output measure.

2.3 Review of the Demand Specification

In this section, we shall review the main characteristics of the various demand studies cited earlier to attempt to draw some conclusions with respect to the specification of a demand model.

2.3.1 Money Illusion

Taylor (1980)'s specification of the demand for telephone services allows for the possibility of money illusion. Only the earlier models such as Bell Canada (1976) and Dreessen (1977,1978) have maintained this format, all other studies rejecting a priori any money illusion. This will also be the approach adopted here.

2.3.2 Cross-price Elasticities

The only attempts at measuring cross-price elasticities are found in the work of Concordia University and in Dreessen (1977, 1978). Cross-price elasticities were estimated unsuccessfully in Bernstein et al. (1977) and Corbo et al. (1978, 1979). The problem follows from the lack of variability of relative prices, the degree of multicollinearity being very high, and from the aggregate nature of the local services series. Similarly, in the Intra-B.C. model, Piekaar (1980) abandoned Dreessen's (1977, 1978) previous attempts to introduce, in some mileage bands, the price of local services as an explanatory variable. If the price of local services is seen as an access charge, it should be deducted from the income variable and not be introduced as a price variable (Bernstein, 1980). Toll calls over other mileage bands are not substitutes, and their price level would act only through the income constraint. In other words, cross-price elasticities between mileage bands should not be a problem to worry about. Cross price elasticities between types of call and between different periods of the day or days of the week are still outstanding problems.

2.3.3 The Dynamic Structure of Demand

The most general linear demand model, in terms of its dynamic structure would be the transfer function:

$$(1)\quad r(B)q_t = B^s\, s(B)p_t + B^u u(B)y_t + B^v v(B)x_t + e_t$$

$$\emptyset(B)e_t = \theta(B)a_t$$

where r(B), s(B), u(B) and v(B) are proper rational functions in B which is itself the lag operator, i.e. such that $Bz_t = Z_{t-1}$,

s, u and v are non-negative scalars which indicate the dead time, $\emptyset(B)$ and $\theta(B)$ are proper polynomials in B and a_t is $N(o,\sigma)$,

q_t, p_t, y_t and x_t denote the quantity demanded, the price of the service, the income and other exogeneous variables, after proper deflation and transformation, as required.

This general model, without transformation of the variables and with $r(B) = (1-r_1B)$, $s(B) = s_o$, $u(B) = u_o$, $v(B) = 0$ and $\emptyset(B) = \theta(B) = 1$ is the Houthaker-Taylor flow adjustment model.

As noted earlier, the standard application in modelling the demand for telephone services is the DL model, which implies that the variables are expressed as logarithms. Then we have the habit formation model; this model was used by Piekaar (1980). If in addition to the habit formation hypothesis, one also assumes that $\emptyset(B) = (1-\emptyset_1 B)$ while $\Theta(B) = 1$, i.e. if one introduces a correction for autocorrelation, then we obtain the model adopted in Corbo et al. (1978) and Fuss and Waverman (1981). The most extensive study of $\emptyset(B)$ can be found in Corbo et al. (1979), in which even $\emptyset(B) = (1-\emptyset_1 B-\emptyset_2 B^2-\emptyset_3 B^3)$ was investigated. However as all of these tests in Corbo et al. (1979) are applied to regressions which also contained cross-price elasticities and as it is unlikely that we can disentangle those elasticities from one another, the utility of these tests is limited. On the other hand, the attempt to go beyond a first degree polynomial in the specification of $\emptyset(B)$ is welcomed since its higher degree polynomials introduce the possibility of complex roots corresponding to cyclical movements.

Economic theory has nothing to tell us as to the proper dynamic structure of the model, and one must turn toward time series analysis. Box and Jenkins (1970) present a methodology to identify a transfer function, however, their methodology cannot be applied, at this stage, to our problem because the series are too short. Furthermore it has also been noted that different models may yield very similar summary statistics (Granger and Newbold, 1978). In this context, it seems wise to follow Box and Jenkins' parsimony principle, i.e. to select the simplest of the models which can reasonably be entertained. The testing, as indicated above, will remain rather ad hoc as long as we do not have longer time series.

2.3.4 The Seasonality of Demand

As noted earlier when referring to the construction of price indexes, seasonality affects the demand for message toll services. Again the time series framework presented in the previous section can accommodate this new dimension without any problem. The problem, however, is that the requirement on data is too much greater, hence, the practical application of standard time series procedure will have to wait for a few years, when, barring major structural changes, we will have sufficiently long series!

The seasonality question affects only the industry studies. There, the Intra-B.C. model is based on seasonally adjusted data, the seasonal adjustment procedure being the Bureau of the Census X-11, while Bell Canada (1980) utilizes seasonal dummies for the intercept. Cleveland (1972) has shown that the X-11 program can be approximated by a seasonal multiplicative autoregressive-integrated-moving average (ARIMA) model. It can also be shown that the use of dummy variables can be analyzed from the point of view of ARIMA models with common roots which were studied by Abraham and Box (1979). Courchesne, Fontenay and Poirier (1980) have developed a general analytical framework within which the use of the X-11 and the use of dummy variables both can be evaluated within the general ARIMA specification. They show that it is an empirical matter which of the dummy variables approach and the use of variables adjusted by the X-11 dominates. Hence, once again, as with the previous sections, we cannot derive a general rule.

Finally, even though it cannot be said exactly how many degrees of freedom are lost through a prior adjustment by the X-11 seasonal adjustment program, if one uses its ARIMA approximation as given in Cleveland (1972) or Cleveland and Tiao (1976), one can obtain a reasonable estimate. This correction was not done in the Intra B.C. model which treats the seasonally adjusted data as raw series.

2.4 Fuss and Waverman (1981) Demand for Toll Calls by Distance and Length of Call

As noted earlier, most approaches adopted in the demand for message toll services are tailored to the data base available to the author. A particularly interesting example is one of Fuss and Waverman (1981)'s models which they tailored around a 1977 Quebec interrogatory which made public the number of calls by duration and by mileage band for Bell Canada. Even though the econometric estimation, by the authors' own account, was unsuccessful, it is worthwhile to present their model.

Bell Canada's message toll tariffs are two-part tariffs, the first minute being more expensive than subsequent minutes which are all always priced at the same rate, and they attempt to tailor their model to that feature. The demand is differentiated in terms of mileage and in terms of duration; while mileage is used to index the demand curves and won't be considered further, the duration is used to specify different forms depending on the number of minutes.

For calls of one minute duration or less, it is assumed that the quantity, i.e. the number of calls, x_1, is a function of the price of one minute calls, p_1, and the expenditures on all toll calls, E:

(2) $x_1 = D^1 (p_1, E)$

For calls of j minutes duration, j > 1, the price of the jth minute (which is independent of j for j > 1), p_2, the expenditure on all toll calls and the access charge for each of the calls lasting j minutes, which will be denoted by $E\ j=(p_1-p_2)xj$, are assumed to be arguments of the demand function, such that

(3) $x_j = D^j \ (P_2, E^*_j, E) \qquad j = 2,3,\ldots 6$

These models were estimated, using the following specification of the demand functional form:

(4) $S_{i,j,t} = \alpha_{i,j} + \beta_i \ln P_{i,j,t} + \gamma_i \ln E^*_{i,j,t} + \delta_i \ln E_t$

Where $S_{i,j}$ is the share of toll call expenditures in mileage band i of duration j

$P_{i,j}$ is the price of the last minute for a call in mileage band i lasting j minutes, i.e. if $j = 1$, $P_{i,j}=P_{i,1}$ and if $j \neq 1$, $P_{i,j}=P_{i2}$,

(5) $E^*_{i,j} = (P_1 - P_2) \ x_{i,j}$

No mention is made by the authors, in their estimation, of the fact that $\sum_{i,j} Si,j,t = 1$ implies the constraints

$$\sum_{i,j} \alpha_{i,j} = 1, \sum_{i} \beta_{i} = \sum_{i} \gamma_{i} = \sum_{i} \delta_{i} = 0$$

However, there are numerous other reasons why one would expect the problems they faced in their estimation of this model. First of all, the disaggregation of the data is misleading; even though they consider seventeen distinct mileage bands, the data are nevertheless aggregated over (i) types of customers, i.e. business and residential, (ii) types of call, i.e. DDD, SOH, and P-P, (iii) carriers involved, i.e. intra-Bell, or adjacent, or TCTS, and finally (iv) time of day - day of week. Experience with the Intra-B.C. model leads us to believe that (ii) is crucial[6], while other things equal, the fact that the call is intra, adjacent, TCTS or US will generally affect its rate. In addition, the methodology to derive the price indexes is not presented.

The specification says nothing about the determinant of the total expenditures on toll calls, E, and it is not fully consistent with the conceptual argumentation for the analytical model. In the latter, it is stated, i.e. that the substitutability between calls is fundamental to the argumentation, yet, in the application, each duration is taken by itself independently of the others.

"...The demand for one minute calls can decrease when the price beyond the first block falls, since a longer call is a substitute for one minute call."

Finally, it is hard to accept the hypothesis that the demand for each duration is inversely related to price; it does not seem to be far fetched to imagine the demand for 1 minute calls, and even, possibly, 2 minute calls to be upward sloping.

2.5 Bell Canada's Intra Model

In considering the modelling of the demand for message toll services, one must repeatedly refer to Bell Canada's path-breaking study, as we have done already and as we will have to do further on in the paper. In this section the intent is to complement our comments by a review of Bell Canada's methodology. As illustrated by table 1, below, the B.C.-Alberta D.L. model considered here is in its general form very similar to the Bell Canada model.

The essential differences are the externality variable which we reject on the grounds that the estimate of its coefficient cannot be justified on economic grounds, (this is discussed in detail in the next section), the level of aggregation across time of day, rate groups and subscribers. The level of aggregation across mileage bands found in Bell's model does not appear justifiable in view of the observed differences in estimated elasticities at lower aggregation levels. The level of aggregation across subscribers raises just as many problems. For the business subscribers, telephone services are one input in their production process, and in their decision process, they will select a level of demand in terms of the price of that service, that of other inputs such as labour, capital and the level of demand. On the other hand the residential customer should consider the income constraint, the price of the service and the price of other goods and services he may demand. It follows that the price of the service is the only variable which is common to both decision processes. The ambiguity in the choice of exogenous variables in Bell's model follows from this problem.

Table 1

British Columbia Telephone Company

Comparison of the B.C.-Alberta D.L. Model
And the Bell Canada Intra Model

B.C.-Alberta D.L. Model	Model	Bell Canada Intra Model
Message-minutes[1] for residential/DDD/day disaggregated into seven mileage bands per resident	$\ln Q_t$ $=$	Deflated revenue for business-residential/DDD/Day-Eve-Night disaggregated into two mileage bands
	a	
Implicit chained price index	$+$ $b \ln P_t$	Laspeyres chained price index
Implicit provincial personal disposable income[3]	$+$ $c \ln y_t$ $-$	Average quarterly number of employed persons 15 years and over, Quebec & Ontario [2]
CPI - Vancouver	$(b+c) \ln CD_t$ $+$	GNE deflator [5]
Rejected on the ground that, due to multi-collinearity, the estimated coefficients do not make any economic sense	$d \ln EXT_t$ $+$	Sum of business and residential main telephones in service[4] replaced in 1981
Seasonal dummy variables	e SEAS $+$	Seasonal dummy variables
Strike dummies, including the postal strikes (1975 Q4 and 1978 Q4)	f STRIKES $+$	Not included
Not included	g DAY	Correction for the number of weekdays in the quarter

Notes:

(1) Since there is no proper disaggregation by rate groupings available, this measure is the best guess. Nevertheless, it fails to give weights to message-minutes which are over greater distances; this problem is minimized by considering seven mileage categories; and DDD only for residential customers during the day.

(2) Bell Canada considers the aggregate over business and residence; as such there does not exist a "proper" income variable.

(3) There is a certain ambiguity in the income variable which is not found in the Bell model, namely since B.C. subscribers can also receive calls from Alberta, Alberta's personal disposable income could be expected to have some impact on calls originated in B.C.

(4) In addition to the observations made in this paper, it should be noted that business main telephones, while technically well defined by any one carrier, is not an unambiguous concept from the demand for message toll services.

(5) The GNE deflator is appropriate only for business customers. Even then, it is not proper here as it accounts for price fluctuations across the whole of Canada, i.e. it reflects price changes which do not intervene in the decision process of an Ontario or Quebec business subscriber.

However these features are not unique to the Bell model. What is more fundamental in that model is the thoroughness of the statistical testing and the implied philosophy. A clear set of criteria are set to evaluate all equations; these cover (i) the use of a residual plot, and (ii) of a normal probability plot, (iii) the F statistic at 1% significance level, (iv) both the R^2 and the R^2 together with the standard error of regression, (v) the t-test at 5% level, (vi) the D.W. statistics and Box-Jenkins' procedure (vii) the Anderson-Darling test of normality, (viii) the Goldfeld-Quandt method to test for heteroscedasticity at the 5% level, (ix) Chow's tests for stability, (x) Klein's procedure to evaluate multicollinearity, (xi) the General Linear hypothesis, and (xii) the appropriateness of deflating by a general price indicator. This is complemented by an ex-post forecast analysis of the results.

The philosophy is outlined at the very beginning of Bell Canada (1980):

> "Because econometric models produce statistically optimal estimates only when the models satisfy certain statistical assumptions, failure of the models to satisfy all of these assumptions can lead to erroneous conclusions." (p.1)

While this position is useful, it is meaningless by itself since it cannot be used to evaluate spurious relations. There are two possible approachs to go around that problem. The first approach is statistical and again suffers from the shortcomings of remaining within the statistical discourse, however it would help us cope with the problem. That solution consists of testing for Grouper Causality. It cannot be applied empirically here for lack of data. The second approach consists of first deriving the bounds to meaningful results from economic analysis. It is hinted at for instance (pp. 5 and 8) when it is stated that "the explanatory variables included in the models were selected on the basis of market characteristics and conventional economic principles..." and "the estimated coefficients of the economic variables must have a plausible sign based on economic principles".

However at no time does the study consider the economic implication of the model beyond those few generalities. It follows that, even though the statistical testing is meritorious, it is also vacuous.

Certainly, economic principles cannot be used in the form of a statistical test such as the t-test. To wit, even though it is known that the price elasticity should be negative, it is also known that there exists situations in which it will be positive. All this means is that econometrics is by necessity as much an art as a science and that its statistical components cannot meaningfully be used mechanically.

3. Network Externalities

Taylor (1980) considers two forms of externalities when he notes that:

> "A completed telephone call requires the participation of a second party, and the utility of this party is accordingly affected... an externality is thereby created... the externality is a call (or use) externality".

and that:

> "Connection of a new subscriber confers a benefit on existing subscribers because the number of telephone that can be reached is increased. In this case, the externality is an access (or system) externality".

The first form of externality is fundamental to any form of communication and in particular to telephony. It is an externality since, presumably, the second party would be willing to pay a price to receive (or, maybe, not to receive) that call. As, in practice, the cost can usually be shared by taking turns, and in the case of a connection which is charged in terms of usage, this form of externality should not be relevant as long as both parties are facing the same price.

The comparability of income between Alberta and British Columbia and the fact that the rates are the same in both directions should minimize the impact of the form of externality.[7]

Now if one additional subscriber joins the network, it is contended that every subscriber will benefit since every subscriber can now reach that new subscriber. This form of externality should have an enormous if not explosive impact on the network since the nth subscriber increases the number of possible originating calls by 2 (n-1), that is the number of possible connections by (n-1). This form of externality, however, depends crucially upon the existence of a potential call externality since it is necessary that the other subscribers increase their usage of the network or at least increase their option demand to gain access to the network to create an externality.[8] In this context, it is wise to distinguish between two types of expansion in the number of subscribers depending upon whether (i) the population does not change but the penetration rate increases, i.e. the increase in n, the number of subscribers, is solely due to individuals who did not have the telephone previously and who are now getting it, or (ii) the population increases at the same rate as the number of subscribers, leaving the penetration rate constant. Whereas type (i) expansion might characterize networks such as the French one, it seems clear that, at the very least through the seventies, it is type (ii) expansion which characterizes the Canadian situation.[9] While the access externality could be expected to be significant given type (i)[10], this should not be true of type (ii) expansion.

It is recognized that the issue is empirical but it is suggested that the econometric approach will not provide the proper test to sustain or refute our contention. First of all, we begin by suggesting that the proportion of inward movements or net gains caused by such effects as "keeping up with the Jones's" is, for all practical purpose, null. We also suggest that new subscribers do not consider, in their decision to join the network, whether others are subscribing. Rather we assume, that they take it for granted that almost everyone is a subscriber. In other words we suggest that access is not affected, in a Canadian context, by the so-called network externality. Our hypothesis could be investigated through a proper survey of new subscribers to establish their rationale for joining the network. To look at the access externality impact on usage, we suggest that a proxy might consist in considering calls per station in two communities with the same penetration rate, relatively similar income and socio-demographic characteristics except for their populations. (We are also making the assumption that usage behaviour doesn't vary much across Canada). Ideally, these two communities should be in the same rate group to avoid a price differential effect. However, it is suggested that as long as it is not too widely different communities, in terms of culture, say Edmonton and Quebec, which are selected, even differences in rate group, should not affect significantly calling patterns. For instance, it is suggested that on a main station basis Victoria would not differ all that much from Vancouver. This hypothesis is based on the contention that the subset of subscribers any one subscriber is likely to call with non-zero probability is extremely small compared to the set itself. Hence a change in the set in the form of a new subscriber could not affect the calling pattern of but a very small subset of subscribers with a non-zero probability; the possibility to reach millions of subscribers by accessing the network is of no real relevance to me as a new subscriber since I am concerned only with the few I am likely to ever reach. Furthermore, even though some old subscribers will reach me with a non-zero probability, it is contended that this will be dominated by substitution in their calling patterns. The argument presented here could be re-phrased in terms of time allocation by consumers who maximize their utility function; in a near saturated market the constraint in using the telephone is time, while in a market with low penetration, the opportunity cost to an increased penetration is likely to be high.

It could be contended that what one measure through such an externality is a change of behaviour, the new subscriber being different from the old. Once again, however, this would have to be established by direct comparison since it appears extremely unlikely to be very significant. In the case of Bell Canada (1980), if this were the interpretation given, one would have to accept that the new subscribers consume in the order of 133 to 145% more message toll services than established customers within Bell territories. If an independent investigation were to establish that this were true, the interpretation of the estimated coefficient of the price variable as an elasticity measure would still be invalid; the regression coefficient is obtained from the reduced form of the aggregation of two distinct subscriber populations, the old and the new subscribers, while the composition of that population is changed.

Finally, it should be noted that, in any case, the number of subscribers that can be reached will vary significantly and simultaneously with the mileage band and the exchange the call originates from. Unless one either takes a point to point approach, or considers the demand originating at the exchange level, it is clear that first the number of subscribers cannot be specified and that it will always differ from the total number of subscribers. Hence, even though the number of subscribers that can be reached could be roughly approximated by the total number of subscribers as a first approximation if the intent is only to deflate the output variables, its interpretation as an externality variable is invalid if it is used as an explanatory variable.

As an illustration of our argument that the inclusion of a network externality variable might produce misleading results, we have estimated the following equation:

$$(6)\quad \ln Q = a_0 + a_1 \ln p + a_2 \ln y + a_3 T + a_4 S1 + a_5 S2 + a_6 S3 + a_7 BCTS + a_8 OKTS + u$$

Where Q is the number of call minutes, Y is personal disposable income in B.C., T is the total number of residence main stations in B.C., S1 - S3 are seasonal dummies and PS, BCTS and OKTS are postal strike, B.C. Tel strike and OKT strike dummies respectively.[11] The variable T is the market size variable to capture the impact of network externality. The resulting estimates are shown in Table 2.

It can be seen that the equation suffers from three major defects. First, 4 of the 7 income coefficients and one price coefficient have the wrong signs. Second, only 3 of 14 elasticity coefficients are significant. And finally, the magnitudes of the market size coefficients are implausibly high; a 1% increase in the number of subscribers will cause a more than 3% increase in the number of toll calls in 6 of 7 mileage bands. This last finding is surprising in terms of expectation but not in terms of past experience. It also concurs with Dreessen's 1981 remarks regarding network externality; namely "failure to specify some kind of per capita model for toll demand may lead to econometric nonsense".

The number of telephones in Alberta instead of the number of residence main stations has also been employed in the estimation as it indicates the market that a B.C. caller can reach. Briefly, the results (not presented here) are:

(1) four of the price coefficients obtain the wrong sign, (2) all income coefficients obtain the right sign with four significant at the 5% level, (3) all market size coefficients obtain the right sign with five significant at the 5% level and (4) the values of the significant market size variable now range from 1.62 to 1.79. The last finding is a considerable improvement over that of Tables 3.7 but, nevertheless, the values of the market size coefficients are still quite high.

Table 2

Price and Income Elasticities
Double Log Model: Dependent Variable = ln Q

	A	B	C	D	E	F	G
CONST	-0.39	-16.36*	-21.77*	-28.44*	-24.42*	-13.53*	-18.88*
	(0.03)	(14.15)	(12.34)	(19.50)	(10.32)	(3.87)	(5.05)
ln P	-2.18°	- 0.05	- 0.14	- 0.03	- 0.30	- 1.27*	- 0.76°
	(2.03)	(0.29)	(0.61)	(0.09)	(1.03)	(3.24)	(1.82)
ln Y	0.95	- 0.21	0.20	0.35	- 0.71	- 0.89	- 0.79
	(0.51)	(0.32)	(0.30)	(0.36)	(0.88)	(0.86)	(0.71)
S1	-0.06	- 0.07*	- 0.05	- 0.14*	- 0.14*	- 0.19*	- 0.19*
	(0.74)	(2.28)	(1.48)	(3.59)	(5.09)	(3.71)	(3.40)
S2	-0.09	0.02	0.05	- 0.14*	- 0.14*	- 0.14*	- 0.13*
	(0.95)	(0.72)	(1.62)	(3.27)	(3.79)	(3.05)	(2.55)
S3	-0.05	- 0.00	0.06*	- 0.06	- 0.10*	- 0.08°	- 0.11*
	(0.74)	(0.04)	(2.14)	(1.61)	(2.84)	(1.89)	(2.26)
PS	0.30	- 0.12	0.27*	0.18	0.19	0.46*	0.51*
	(0.98)	(1.01)	(2.25)	(1.07)	(1.35)	(2.52)	(2.60)
OKTS	-0.15	- 0.00	0.09	- 0.18°	- 0.07	- 0.05	- 0.05
	(1.35)	(0.25)	(0.57)	(0.14)	(0.47)	(0.83)	(0.47)
BCTS	-0.15	- 0.00	0.09	- 0.18°	- 0.07	- 0.05	- 0.05
	(0.76)	(0.05)	(1.32)	(1.81)	(0.81)	(0.45)	(0.44)
ln T	-1.62	3.17*	3.65*	4.57*	5.20*	3.76*	4.47*
	(0.48)	(4.63)	(5.08)	(4.48)	(5.93)	(3.34)	(3.65)
F(9,28)	31.99	121.19	211.67	160.92	219.73	124.40	107.64
R	0.941	0.984	0.991	0.988	0.991	0.984	0.982
d.w.	1.72	2.03	2.43	1.33	0.85	1.02	1.36

* significant at 5%
° significant at 10%

4. A Message Toll Services Demand Model

It was pointed out earlier that recent Canadian results confirm the result that the own price elasticity of demand is not constant with respect to distance. In addition to Fuss and Waverman (1981)'s conclusion to that effect, we can also cite Breslaw's (1980) disaggregation of Bell Canada's message toll services. However, the most convincing evidences are provided by the Bell Canada (1980) model and by the Intra-B.C. model.

The DDD business and residential (0-100 miles) service and the (101+ miles) service yield very similar results on the whole, except for the income elasticity which appears to decrease as the distance increases: from .38 to .24, and the own price elasticity which almost doubles: from -.18 to -.32. As such, one has to be careful when interpreting the income elasticity variable adopted by Bell Canada which is only a proxy for income; in fact it is hard to understand why such a variable (the average number of employed persons in Quebec and Ontario) was adopted by Bell Canada when, consistently, in recent rate applications, Bell Canada has also been presenting in the form of an exhibit to support their application the graph of the year-to-year percent increase in toll messages shown next to that of the year-to-year percent increase in the GNE measured in constant dollars (Bell Canada, Exhibit B-81-220). In addition, at least one of these elasticity estimates is not significantly different from zero.

The Intra-B.C. model price elasticities, given in table A.1 and A.2 in the Appendix, once again confirm that, especially with respect to DDD-type calls, the elasticity increases rather systematically with distance. In addition, when we observe the income elasticity (Table A.2), we observe a tendency for the elasticity to decrease with distance which is generally consistent with that observed in Bell Canada (1980).

We suggest that such evidences are sufficient to raise serious questions as to the applicability of the DL model to the analysis of demand for message toll services. We will further use this result to hypothesize such an observation implies that the own price elasticity and the income elasticity of message toll services should, at least, be allowed to vary systematically with price and income.

A telephone conversation enables two parties to communicate with one another. There are two features which are unique to modern telecommunications: first, the communication is nearly instantaneous, second the quality of the communication is approximately independent of distance. On theoretical grounds one cannot, however, assume that distance is not an argument in the utility function of the consumer since the opportunity cost to the consumer, to the extent it can be defined, may be expected to increase very rapidly with distance ("may" since it won't as long as the alternative is the mail while it will in all other cases since transportation will be involved). Nevertheless since there is no readily available close substitute, any form of transportation being time consuming and, with the exception of the mail, very costly, we shall nevertheless assume that distance is not an argument on the consumer's utility function.[12] Evidently our assumption also assumes that all the functions fulfilled by the telephone are not distant specific; this is clearly not the case when we consider the complete set of distances since such services as emergency services are relevant solely over very short distances. Since we restrict our attention to message toll services, it is felt that it is a very weak assumption. Given such an hypothesis, it is possible to pool observations across mileage bands in the same demand model. To be consistent with existing observations, this model must be such as to account for the higher elasticity associated with calls over longer distances. To account for this higher elasticity, while at the same time recognizing the rejection of distance as the determinant of this elasticity variation, we note that, empirically, distance is positively correlated with the price of a message-minute. Hence, we assume that the observed correlation in elasticity with distance is the result of a utility function which is such as to associate a higher elasticity with a higher price. Evidently, we do not contend that the price level is the cause of the elasticity, rather that the observed correlation is the result of the form of the utility function.

While Breslaw and Smith (1981) had noted that:

> "Problems including multicollinearity and a small sample size effectively preclude the accurate estimation of cross-elasticity terms or terms which would allow the own-elasticities to vary with price and income".

it can be expected that the pooling of mileage band observations with time series will create sufficient variability in a sufficiently increased sample size to obtain accurate estimates of changes in own price elasticities with respect to price and income. Cross-elasticity terms shall be ignored on the grounds that calls over different distances are not proper substitutes for one another.

As the data which we are using in this paper are not point-to-point, it is not possible to clearly identify the population within B.C. which can make the calls and the population within Alberta which can be reached. Furthermore, it is likely that there will be considerable variations within as well as between mileage bands since, for the smallest mileage bands, B.C. Tel subscribers living in Vancouver or Victoria are excluded. In fact, mileage bands F and G include the toll calls from Vancouver and/or Victoria to Calgary and/or Edmonton respectively. Even if the number of message-minutes are indexed in terms of the B.C. Tel subscriber population and the Alberta subscriber population, a B.C. Tel subscriber selected at random is more likely to make a call in mileage bands F and G.

This will be accounted for through dummy variables which let the levels of the demand curve vary with the mileage bands. Furthermore, to account for variation in demand characteristics between various point-to-point combinations, variations which could be due to demographic or geographic characteristics, we begin by specifying the demand function as a general function such that

$$(7)\quad q_{i,t} = F(P_{i,t}, P_t, Y_t, D_i),\ i = 1, 2, \ldots, 7$$

where $q_{i,t}$ is the number of message-minutes in mileage band i, period t, by subscribers in B.C. Tel, to subscribers in Alberta, $P_{i,t}$ is the price index of a message-minute in mileage band i, period t, P_t is the general price level in period t, Y_t is the per capita personal disposable income in period t, and D_i is a vector of dummy variables, (δ_j), such that

$\delta_{i,1} = 0$ for all i, and

$$\delta_{i,j} = \begin{cases} 1 & \text{for } i = j \\ 0 & i \neq j \end{cases} \qquad j = 2, 3, \ldots., 7$$

It can further be assumed that there exists a flexible functional form, defined by Diewert (1973), which "contains precisely the number of parameters needed to provide a second order approximation to an arbitrary twice differentiable ... function satisfying the appropriate regularity conditions...", where F is that arbitrary function. A possible approach consists of considering it as a second-order Taylor's series approximation (Blackorby, Primont and Russell, 1978).

Since accumulated experience indicates that the DL model gives good fit, it is reasonable to look for a flexible form which is closely related to the DL model. As the DL model can be seen as a first-order Taylor's series approximation in the log to any appropriate demand function, the logical extension would be the second order approximation in the log.

Assuming no money illusions, i.e. zero homogeneity with respect to all prices, the demand function can be written

(8) $q_{i,t} = f\ (\rho_{i,t},\ y_t,\ D_i)$

where $\rho_{i,t}$ and y_t correspond respectively to $(P_{i,t}/P_t)$ and (y_t/P_t).

Following Jorgenson and Nishimizn (1978), the translog demand function will be

$$(9)\quad \ln q_{i,t} = \sum_{i=1}^{7} \alpha_{i,j}\, \delta_{i,j} + \alpha_p \ln \rho_{i,t} + \alpha_y \ln y_t + \sum_{j=1}^{7} \beta_{p,j}\, \delta_j \ln \rho_{i,t} + \sum_{j=1}^{7} \beta_{y,j}\, \delta_j \ln y_t + \beta_{p,y} \ln \rho_{i,t} \ln y_t + \frac{1}{2} \beta_{p,p} (\ln \rho_{i,t})^2 + \frac{1}{2} \beta_{y,y} (\ln y_t)^2$$

where $\alpha_{0,0}$ and $\sum_{j=1}^{7} \beta_{\delta,j}\, \delta_j^2$ have been omitted since δ_j^2 cannot be differentiated from δ_j, and where $\delta_{i,j}$ is as previously specified.

This form of the translog reduces to the DL, applied independently a mileage band at a time, whenever $\beta_{p,y} = \beta_{p,p} = \beta_{y,y} = 0$(13) since then

$$(10)\quad \ln q_{i,t} = \gamma_{0,i} + \gamma_{p,i} \ln \rho_{i,t} + \gamma_{y,i} \ln y_t$$

where $\gamma_{o,i} = \alpha_{o,i}$

$\gamma_{p,i} = \alpha_p + \beta_{p,i}$

$\gamma_{y,i} = \alpha_y + \beta_{y,i}$

The problem we face with this general form is that raised by Breslaw and Smith, even though now the elasticity can vary systematically with the price and the income:

$$\text{(11)} \quad \frac{d \ln q_{i,t}}{d \ln p_{i,t}} = (\alpha_p + \beta'_{p,i}) + \beta_{p,y} \ln y_t + \beta_{p,p} \ln p_{i,t}$$

$i = 1, 2, \ldots,$

$$\frac{d \ln q_{i,t}}{d \ln y_t} = (\alpha_y + \beta'_{y,i}) + \beta_{p,y} \ln p_{i,t} + \beta_{y,y} \ln y_t$$

The number of observations on the demand function has now been multiplied by seven, while the same was done to the number of parameters to which three new parameters are added. Part of our intention in developing this model was based on the desire to retain as many degrees of freedom. In this context it seems plausible to further assume that the geographic and demographic characteristics of demand only affect the level of demand without having any impact on the elasticities.

The demand function becomes

$$\text{(12)} \quad \begin{aligned} \ln q_{i,t} = {} & \alpha_{0,1} + \sum_{j=2}^{7} \alpha_{0,j} \delta_{ij} + \alpha_p \ln p_{i,t} + \alpha_y \ln y_t \\ & + \beta_{p,y} \ln p_{i,t} \ln y_t + \frac{1}{2} \beta_{p,p} (\ln p_{i,t})^2 \\ & + \frac{1}{2} \beta_{y,y} (\ln y_t)^2 \end{aligned}$$

which yields as own-price elasticity

$$\text{(13)} \quad \frac{d \ln q_{i,t}}{d \ln p_{i,t}} = \alpha_p + \beta_{p,y} \ln y_t + \beta_{pp} \ln p_{i,t}$$

and, as own-income elasticity

$$\text{(14)} \quad \frac{d \ln q_{i,t}}{d \ln y_t} = \alpha_y + \beta_{p,y} \ln p_{i,t} + \beta_{y,y} \ln y_t$$

Furthermore, to the extent that the observed differences in estimated elasticities are indeed the result of differences in the price level prevailing in various groupings, in Bell Canada (1980) and Piekaar (1980), then the following hypothesis can be entertained:

$$\beta_{p.p} > 0$$

i.e. that the own-price elasticity decreases as the price level decreases. This hypothesis is interesting since it enables us to model and test an empirical observation that might be interpreted by some as a structural change, namely the fact that patterns of usage of message toll services over longer distances appear to have evolved in recent years. An alternative justification which might be offered is that younger subscribers have an inherently different demand function; such a hypothesis could only be tested if (i) one could define a younger subscriber, (ii) and one could relate chronologically a sample of "younger subscribers" and one of "non-younger subscribers".

It can be expected that, as the price of the service decreases, the quantity demanded, while it may increase, will tend toward some maximum rather than increase to infinity. This expectation can be based on the observation that, at the local level, even though additional calls are free, the number of calls for any subscriber is finite. It can conceptually be justified in terms of the opportunity cost of a call to a subscriber, the opportunity cost of a local call being measured in terms of other uses of his leisure time, in a Becher-type analysis.[14] One way to describe the service to the subscriber is to suggest that, as the price goes down, the subscriber considers the service progressively less as a luxury and progressively more as a necessity. As there are no close complements or substitutes, it would seem reasonable that, in the price effect, the income effect dominates. Then it would seem reasonable to assume that , if the own-price elasticity decreases as the price decreases, it is mostly because successively lower price levels create successive income effects which have decreasing impacts of the subscriber's own-price elasticity. In other words we may entertain the hypothesis that

$$\beta_{p.y} < 0$$

All of this leads us to suggest that on Engel's curve it is appropriate to describe the demand for message toll services, hence that

$$\beta_{y,y} < 0$$

This hypothesis is consistent with Bell (1980).

Strictly speaking, these hypotheses, together with the independence of the income and price elasticities with respect to the mileage band and the DL model can be tested since they are all nested in the general translog form. However the lack of price variability within any mileage band together with the shortness of the time series make it unlikely that the test would be meaningful.

Empirical Results

5.1 Data

In this section, the estimated results of the proposed translog demand model are presented. These results are compared with estimates from the double log model to highlight and quantify the differences between the two specifications. For comparability, price and income elasticities from different specifications are estimated using a common data base. As indicated in Section I, the market segment being studied is B.C. to Alberta residence weekday daytime dialed direct traffic. Based on 1979 data, this market segment represents approximately 9.3% and 10.3% of total B.C. to Alberta call minutes and revenue respectively. The data used consist of a sample of calls and corresponding dollars that are compiled on a monthly basis. These calls are regrouped into 7 mileage bands which are listed in Column A of Table 4 with corresponding rate steps in the B.C. to Alberta tariff listed in Column B. Quarterly observations for 1973 - 1979 are used in the estimation. This period covers one rate change and is chosen because the conversion to DDD was not completed until late 1972. Prior to 1973, some operator handled calls were wrongly classified as DDD equivalent. The choice of 1973 as the start year avoids this problem.

In Tables 5 and 6, the major characteristics of this segment are presented. Table 5 contains a brief description for each mileage band of the major routes, originating and terminating points. Table 6 shows the shares of each mileage band in total revenue and call minutes for this market segment. As indicated, mileage bands F and G together represents approximately 60% of the traffic. This is to be expected since these two bands include the two largest centres in each province; Vancouver and Victoria in B.C. and Edmonton and Calgary in Alberta.

Table 4

Definition of Mileage Bands

A		B	
Data Base		Tariff	
Mileage Band	Mileage	Rate Step	Mileage
A	0 - 20	102	0 - 20
B	21 - 80	103	21 - 36
		104	37 - 56
		105	57 - 80
C	81 - 180	106	81 - 110
		107	111 - 144
		108	145 - 180
D	181 - 290	109	181 - 228
		110	229 - 290
E	291 - 400	111	291 - 400
F	401 - 500	112	over 400
G	500 +		

Notes:

1. Prior to August 1975, rate step 102 was subdivided into 2 steps - 1 to 8 and 9 to 20.

TABLE 5

B.C. - ALBERTA LONG DISTANCE CALLING MARKET CHARACTERISTICS

MILEAGE BAND	RATE STEPS	MAJOR CHARACTERISTICS
A	102	All calls originate from B.C./Alberta border with one route (Dawson Creek to Bonanza) accounting for 45% of total traffic.
B	103	All calls originate from B.C./Alberta border, with one route (Sparwood to Blairmore) accounting for 42% of traffic. In fact, approximately 55% of total traffic originates from Sparwood.
B	104	All calls originate from B.C./Alberta border with no dominating route. However, the five routes (out of 47) with the most traffic account for over 40% of total traffic.
B	105	All calls originate from B.C./Alberta border with one route (Dawson Creek to Grande Prairie) accounting for 53% of total traffic.
C	106	All calls originate from B.C./Alberta border with five routes out of 159 accounting for over 60% of traffic.
C	107	Most calls from South-eastern B.C. with one route (Cranbrook to Calgary) accounting for over 36% of traffic and the five routes (out of 213) with the most traffic accounting for over 60% of total traffic.
C	108	Most traffic originates from B.C./Alberta border and south-eastern B.C. Out of 330 routes, five accounted for over 40% of total revenue.

TABLE 5 - (Continued)

MILEAGE BAND	RATE STEPS	MAJOR CHARACTERISTICS
D	109	Most traffic originates from South East of B.C. Out of 673 routes, five accounted for 38% of total revenue. These five routes all terminate at Calgary.
D	110	Most traffic from the Okanagan Valley with five out of 1307 routes accounting for 46% of revenue. These five routes all terminate at Calgary.
E	111	Calls appear to come from all over B.C. with no single route accounting for more than 10% of traffic. However, five out of 3413 accounted for 26% of total revenue. The five routes are from Dawson Creek, Fort St. John, Kelowna and Prince George and all terminating at Edmonton.
F & G	112	There are two mileage bands in this rate step. The major routes are Vancouver and Victoria to Calgary in mileage band F and Vancouver and Victoria to Edmonton in mileage band G. These four routes accounted for over 60% of traffic in this rate step.

Notes:

(1) Based on December 1979 data.

(2) Per cent figures are approximate.

(3) Include all types of calls.

TABLE 6
Mileage Band

B.C. to Alberta Call Minutes and Revenue by Mileage Bands

	A	B	C	D	E	F	G	TOTAL
MINUTES	3518	14802	56112	83172	110160	196258	155210	619232
SHARE	0.6	2.4	9.1	13.4	17.8	31.7	25.1	100%
REVENUE	402	4283	24431	50054	74523	142084	112034	407811
SHARE	0.1	1.1	6.0	12.3	18.3	34.8	27.5	100%

NOTE: (1) Total is approximately 9.3% of total B.C. to Alberta call minutes and 10.3% of total B.C. to Alberta revenue.

(2) Based on 1979 sample.

As discussed in section 4, the dependent variable is minutes of calling per residence main station. The independent variables are own price, per capita income and a vector of dummy variables to account for seasonality and discontinuous shocks such as strikes. The own price variable is measured as revenue per minute of call. This definition is employed for data reasons as the present sampling file does not sample data by rate steps so that actual price per minute of call as given in the rate table can be used.[15] The correspondence between this price and the actual price as given in the rate table is presented in Table 7. Part A shows the prices for the initial and each additional minute of call for customer dialed calls and part B shows the implicit price employed in this study. For each part, the price before and after the August 1975 rate change are presented. As can be inferred from this table, there are at least two problems associated with the use of implicit prices. First, any change in the mix of calls (for example, changes in the relative shares of steps 106, 107 and 108) will show up as a price change even when there is no rate change. And second, multi-part tariffs for some segments are also ignored. In addition, there is also the problem of Christmas and New Year days that fall on a week-day. For those days, there are discounts for DDD calls but our sampling procedure does not allow for that.[16] These are important problems that should be borne in mind when interpreting the estimation results.

The income variable used is per capita personal disposable income in B.C. in real terms. It has the same value for all mileage bands at any given point in time and has not been adjusted for expenditure on telephone services. The effect of the latter is insignificant as the proportion of household expenditure on telephone services is likely to be very small. The lack of a household income measure for each mileage band separately, on the other hand, is more serious. Personal disposable income in B.C. is dominated by incomes in Vancouver and Victoria which may not reflect income movements in other regions of B.C.

The other explanatory variables are all dummy variables to account for seasonality and discontinuous shocks such as strikes in the economy. There are three seasonal dummies and three strike dummies. The value of each strike dummy is determined by the ratio of number of weekdays affected by the strike to the total number of weekdays in the quarter.

TABLE 7

Comparison of Actual and Implicit Price: $ per minute

A: Actual B: Implicit Price

Rate Steps		Before Rate Change	After Rate Change	Mileage Bands	Before Rate Change[2]	After Rate Change[3]
102[4]	1st min	.17[1]	.15	A	0.11	0.12
	add min	.08	.10			
103	1st min	.21	>.22			
	add min	.15				
104	1st min	.25	>.28	B	0.23	0.29
	add min	.21				
105	1st min	.30	>.34			
	add min	.27				
106	1st min	.35	>.40			
	add min	.33				
107	1st min	.40	>.46	C	0.38	0.44
	add min	.39				
108	each min	.45	.52			
109	each min	.50	.58	D	0.53	0.62
110	each min	.55	.64			
111	each min	.60	.70	E	0.59	0.69
112	each min	.65	.75	F	0.64	0.75
				G	0.64	0.75

Notes: (1) Average of two rate steps. Prior to August 1975 rate step 102 was subdivided into two rate steps.

(2) Based on 1974 Q2 Data

(3) Based on 1978 Q2 Data

(4) There is a minimum charge of .20¢ per call.

All dollars values are converted to real terms by deflating by the Vancouver CPI. Precise definitions of these and other variables are given in the Appendix, Table A.3.

5.2 Double Log Specifications

The equation estimated is

$$(15) \quad \ln q = a_\circ + a_1 \ln p + a_2 \ln (y) + a_3 S1 + a_4 S2 + a_5 S3 + a_6 PS + a_7\, BCTS + a_8 OKTS + u$$

where

q = call minutes per residence main station in thousands,

p = own price divided by Vancouver CPI,

y = personal disposable income in B.C. divided by Vancouver CPI,

S1 - S3 = seasonal dummies,

BCTS = BCT strike dummy (1977Q4-1978Q1),

OKTS = OKT strike dummy (1973Q3-1974Q1),

PS = postal strike dummy (1975Q4 and 1978Q4), and

u = random error.

This equation was initially estimated using OLS for each mileage band separately. The results of these regressions are shown in Table 8. All summary statistics appear to be reasonable. The Durbin-Watson statistics indicate 1st order autocorrelation is not a serious problem and the residual plots of all equations appear to be normally distributed with constant variance. As can be seen, 13 of 14 elasticity coefficients had the expected signs, 9 coefficients are significant at the 5% level and 3 at the 10% level. Considering mileage bands B to F, all the income elasticities are in the elastic range though some appear to be unrealistically high and the price elasticities exhibit the expected positive relationship in absolute value with distance.

Table 8

Double Log Model: Dependent Variable = In q

	MB1	MB2	MB3	MB4	MB5	MB6	MB7
CONST	-8.91*	-3.77*	-3.05*	-2.68*	-2.58*	-2.55*	-2.51*
	(7.52)	(8.44)	(6.38)	(5.02)	(5.75)	(7.25)	(6.09)
ln p	-1.54*	-0.23	-0.68°	-0.93°	-1.26*	-1.98*	-1.68*
	(3.46)	(0.98)	(1.94)	(1.83)	(2.66)	(4.91)	(3.58)
ln y	-0.53	2.61*	3.34*	4.07*	3.38*	1.30	2.08*
	(0.38)	(6.79)	15.22	(4.14)	(3.74)	(1.56)	(2.14)
S1	-0.09	-0.01	0.03	-0.06	-0.09	-0.12*	-0.10
	(1.13)	(0.18)	(0.59)	(0.94)	(1.36)	(2.20)	(1.52)
S2	-0.04	-0.02	0.01	-0.17*	-0.17*	-0.16*	-0.14*
	(0.61)	(0.39)	(0.20)	(2.55)	(2.61)	(2.76)	(2.14)
S3	-0.04	0.01	0.07	-0.03	-0.06	-0.06	-0.07
	0.61)	(0.22)	(1.63)	(0.47)	(0.97)	(1.12)	(1.16)
PS	0.28	0.08	0.53*	0.54*	0.55*	0.68*	0.80*
	(0.93)	(0.47)	(2.93)	(2.19)	(2.26)	(3.32)	(3.30)
OKTS	-0.20	-0.05	0.00	-0.05	-0.10	-0.09	-0.00
	(1.51)	(0.72)	(0.01)	(0.49)	(0.90)	(1.01)	(0.03)
BCTS	-0.10	0.01	0.11	-0.14	-0.03	-0.02	0.02
	(0.55)	(0.11)	(0.99)	(0.87)	(0.20)	(0.17)	(0.11)
F(8,28)	21.70	25.93	49.99	44.65	44.94	59.59	43.15
R	0.901	0.916	0.955	0.950	0.950	0.962	0.948
dw	1.69	2.58	2.61	1.86	1.90	1.70	1.97

* Significant at 5%
° Significant at 10%

Because of data and other problems (for example batches of calls missing and calls reported late and therefore excluded from the data sample) which may affect all mileage bands, the equations (15) for all mileage bands were re-estimated jointly with Zellner's Seemingly Unrelated Regression technique. No adjustment for autocorrelation was made as this is not a serious problem as indicated by the OLS results. The result of this procedure is presented in Table 9. Now all the elasticity coefficients have the expected signs and, in general, are more significant (11 of 14 are now significant at the 5% level) than the OLS results. The three insignificant coefficients are the price coefficients for mileage bands B and D and the income coefficient for mileage band A. The reason for the insignificant price coefficients could be explained by the poor quality of data used as prices are measured implicitly, with no adjustment for changing mix of calls. As to the insignificant income coefficient, a possible reason is that personal disposal income for B.C. is not a good variable for the household income of callers in mileage band A. This can be seen from Table 5 which indicates one route (Dawson Creek to Bonanza) accounted for over 45% of total traffic. As we move to a higher mileage band, a greater proportion of the B.C. population is potentially included and the personal disposable income measure used becomes more representative.

Table 9

Double Log Model: Dependent Variable = log q
Zellner's Procedure

	A	B	C	D	E	F	G
CONST							
	-8.43*	-3.91*	-3.02*	-2.37*	-2.40	-2.29*	-2.23*
	(8.12)	(10.94)	(2.40)	(8.01)	(5.43)	(7.00)	(6.81)
ln p	-1.37*	-0.32	-0.66*	-0.63	-1.08*	-1.68*	-1.36*
	(3.49)	(1.61)	(2.40)	(1.53)	(2.97)	(5.09)	(2.65)
ln Y	0.00	2.52*	3.37*	4.60*	3.71*	1.86*	2.70*
	(0.00)	(7.42)	(6.35)	(5.54)	(5.09)	(2.65)	(3.38)
S1	-0.00	-0.01	0.03	-0.06	-0.09	-0.12*	-0.09
	(0.99)	(0.19)	(0.60)	(0.88)	(1.33)	(2.14)	(1.47)
S2	-0.05	-0.01	0.01	-0.19*	-0.18*	-0.17*	-0.16*
	(0.62)	(0.35)	(0.19)	(2.81)	(2.81)	(3.12)	(2.50)
S3	-0.04	0.01	0.07	-0.04	-0.07	-0.07	-0.08
	(0.55)	(0.29)	(1.63)	(0.60)	(1.07)	(1.30)	(1.35)
PS	0.33	0.09	0.52*	0.52*	0.54*	0.66*	0.78*
	(1.10)	(0.56)	(2.94)	(2.11)	(2.22)	(3.26)	(3.25)
OKTS	-0.21	-0.05	-0.00	-0.07	-0.10	-0.10	-0.01
	(1.62)	(0.71)	(0.00)	(0.58)	(0.92)	(1.09)	(0.10)
BCTS	-0.09	0.02	0.11	-0.15	-0.04	-0.03	-0.03
	(0.47)	(0.16)	(0.98)	(0.94)	(0.24)	(0.25)	(0.18)

Degree of freedom = 133
Weighted R-Square = 0.8659

* significant at 5%
° significant at 10%

5.3 Translog Demand Specification

In this section, estimates from the Tranlog demand function proposed in Section 2 are presented. As discussed earlier, the strength of the model is to allow mileage band data to be pooled together; this should introduce wider variations in the explanatory variables and hence partially alleviate the problem of multicollinearity. The number of observations is 196 (7 mileage bands with 28 observations each). The estimated equation is (12) with the following additional variables: three seasonal dummies and three strike dummies (PS, OKTS and BCTS) i.e.

$$\begin{aligned}(17)\quad \ln q_{i,t} &= \alpha_{0,0} + \sum_{j=1}^{6} \alpha_{0,j}\,\delta_j + \alpha_p \ln p_{i,t} + \alpha_y \ln y_t \\ &+ \beta_{p.y} \ln p_{i,t} \ln y_t + \tfrac{1}{2}\beta_{pp}(\ln p_{i,t})^2 + \tfrac{1}{2}\beta_{y.y}(\ln y_t)^2 \\ &+ a_1 S1_t + a_2 S2_t + a_3 S3_t + a_4 OKTS_t \\ &+ a_5 BCTS_t + a_6 PS_t + u_{i,t} \qquad i = 1,2,\ldots\ldots 7 \\ &\qquad\qquad\qquad\qquad\qquad\qquad t = 1,2,\ldots\ldots 28\end{aligned}$$

where the dummy variables are defined as before.

This equation was estimated using OLS and the results are presented in Table 3.10. The use of OLS was justified for the following reasons. First, the results in terms of signs and magnitudes of the estimated coefficients are not significantly different when an alternative error structure is assumed.[17] And second, OLS is unbiased in any event. The results indicate that all price and income coefficients have the expected signs and four of the five coefficients are significant at the 5% and one at the 10% level. Because the translog model is an extension of the double-log model, an F test was also performed to test the composite hypothesis $\beta_{p,y} = \beta_{pp} = \beta_{y,y} = 0$ to find out whether the higher order tems add any explanatory power to the model.[18] The test statistics decisively rejected the null hypothesis at the 1% level.

In Table 11 the estimated price and income elasticities with their corresponding standard errors are presented for each mileage band.

The elasticity estimates are calculated using equations (13) and (14) and the corresponding standard errors provided to indicate the precision of the elasticity estimates. As the elasticities vary systematically with income and price they will change through time; hence they are evaluated at three data points (the mean and the beginning and end of the study period). For rate decisions and planning purposes the most relevant elasticity estimates are those at the end of the period.[19] The results confirm the entertained hypothesis. The price elasticity estimates suggest demand for calls in this market segment is highly price elastic. Both price and income elasticity estimates appear to decline over time, from very elastic to moderately elastic or inelastic. In one case, the shortest mileage band, the income elasticity even turns negative.[20]

Table 10

Trans Log Model = OLS

	B Values	T for H:B=0	PROB |+| > 0
CONSTANT	2.06	8.73	0.00
lnp	-1.85	-5.36	0.00
(lnp)	-0.15	-1.69	0.09
(lnP)(ln$y$)	0.88	1.97	0.05
ln$y$	2.73	3.45	0.00
(ln$y$)	-4.16	-2.14	0.03
BCTS	-0.04	-0.76	0.45
OKTS	-0.06	-1.43	0.15
PS	0.50	5.72	0.00
S1	-0.06	-2.73	0.01
S2	-0.09	-4.11	0.00
S3	-0.02	-0.88	0.38
δ_1	-6.09	-27.28	0.00
δ_2	-3.55	-23.57	0.00
δ_3	-1.73	-17.99	0.00
δ_4	-0.97	-21.36	0.00
δ_5	-0.49	-16.49	0.00
δ_6	0.21	8.42	0.00

Degree of freedom for t statistics = 178

F test(for $\beta_{p,p}=\beta_{p,y}=\beta_{y,y} = 0$) = 26.00

PROB |+| > 0 indicates the probability of getting a larger absolute t if B = 0.

Table 11

Price and Income Elasticities

Trans Log Model: OLS

	A	B	C	D	E	F	G
PRICE ELASTICITIES							
73.Q1	-1.33 (0.22)	-1.22 (0.17)	-1.65 (0.14)	-1.77 (0.20)	-1.80 (0.22)	-1.83 (0.22)	-1.83 (0.20)
MEAN	-1.12 (0.20)	-1.37 (0.14)	-1.50 (0.14)	-1.60 (0.20)	-1.63 (0.20)	-1.65 (0.20)	-1.65 (0.20)
79Q4	-0.94 (0.20)	-1.22 (0.14)	-1.34 (0.14)	-1.43 (0.20)	-1.46 (0.20)	-1.48 (0.20)	-1.48 (0.20)
INCOME ELASTICITIES							
73Q1	2.18 (0.72)	2.73 (0.54)	3.12 (0.47)	3.46 (0.47)	3.55 (0.48)	3.66 (0.50)	3.65 (0.50)
MEAN	0.82 (0.57)	1.55 (0.30)	1.93 (0.26)	2.23 (0.36)	2.32 (0.39)	2.39 (0.41)	2.39 (0.41)
79Q4	-0.33 (0.69)	0.48 (0.54)	0.83 (0.56)	1.10 (0.61)	1.20 (0.63)	1.26 (0.66)	1.25 (0.66)

* Standard errors are given in parentheses.

The general trend of declining income elasticities indicates that telephone calls in this segment are becoming a necessity as households' incomes grow. The declining trend in price elasticities on the other hand indicates consumers tend to be less sensitive to price changes at lower prices. The reason for this is the income effects which offset part of the negative price effects. In fact, examination of the size of the price and income coefficients (Table 10) suggests that income effects appear to dominate over price effects. This is indicated by the relative size of the coefficients $\beta_{p,p}$, $\beta_{p,y}$ and $\beta_{y,y}$. Take price elasticities for example. Both β_{pp} and β_{py} have the right sizes but the size of $\beta_{p,y}$ (.88) gives income much greater impact on elasticity than that of price through $\beta_{p,p}$ (-.15). Combining both the price and income elasticity estimates, the results suggest that as time goes on, if income does rise and new technology lowers cost and price, the service will become more and more of a necessity.

In Table 12, price and income elasticity estimates from the Intra B.C. model, the double log model and the translog model are presented. Comparing the translog model with the double log specification, the income elasticities are generally higher and price elasticity appears to be much more sensitive to mileage in the double log model. A similar picture also emerges when the translog estimates are compared with the Intra B.C. estimates. These differences between the translog model and the other two models are probably due to the inclusion of the mileage band dummies which filter out some of the structural differences between mileage bands.

Table 12

Comparison of Price and Income Elasticities

Mileage Band	Price			Income		
	Intra	Double Log	Trans Log	Intra	Double Log	Trans Log
A	-0.16	-1.37	-1.12	1.69	0.00	0.82
B	-0.35	-0.32	-1.37	1.09	2.52	1.55
C	-0.54	-0.66	-1.50	2.38	3.37	1.93
D	-0.88	-0.63	-1.60	2.51	4.60	2.23
E	-1.60	-1.08	-1.63	2.16	3.71	2.23
F	-2.02	-1.68	-1.65	1.80	1.86	2.39
G	-2.31	-1.36	-1.65	0.72	2.70	2.39

4. Conclusion

In this study, several issues relating to modelling telephone demand are critically discussed, and a more general demand model for toll was proposed. It was used to specify an alternative specification which was empirically tested. The proposed model makes it possible to cope with the empirical observation that price elasticities in absolute value increase with distance. The results indicate that the demand for calls is price elastic and that long distance calls are becoming a necessity as household income grows. The price elasticity findings contradict some widely held beliefs about price elasticities for long distance calls; namely long distance toll calls are price inelastic (see Bell Canada (1980) and Taylor (1980) for example). The policy implication of the above is obvious; instead of raising prices to increase revenue and earnings, the carriers should consider lowering prices, subject to cost considerations of course. However, because of weaknesses of the data, further work is needed to develop a better data base as well as to devise a procedure to choose among competing models.

In addition, the issue of network externality was also examined. It was argued that Taylor's proposed call externality is appropriate only under certain situations and that the inclusion of a market size variable as a predictor in other situations may lead to misleading results. The results have borne out this expectation in that the market size coefficient is unreasonably high. This evidence is consistent with the findings in other studies such as Dreessen (1981). It suggests that, at least in the context of a disaggregated model, an externality variable should not be used as a separate explanatory variable.

Footnotes

[1] See also B.C. Tel's comments on Bernstein's B.C. Tel model (1980).

[2] This is true even of Intra-Bell message toll series, since Intra-Bell implies message toll services in Quebec and Ontario, a territory which includes numerous carriers besides Bell Canada, such as Quebec Tel, Telebec,... This series includes all toll revenues settled with those carriers. Evidently the distortion between originating and settled would still be much smaller than for the two other categories.

[3] Routledge has noted another problem with the chaining process. As toll revenues are highly seasonal and as rate applications do not come at regular intervals, seasonality will affect the weights over time.

As the pattern of calls changes through the seasons, the share of expenditures of the various categories will fluctuate. This fluctuation will be reflected in the price and quantity indexes. Say, for instance, that there are two categories of calls, the short haul calls which are mostly made in the first of two seasons and the long haul calls which are mostly made in the last of two seasons. Say that, as in the last rate case, long haul rates decrease while short haul rates increase, then, depending upon the season in which the price index is computed, we might obtain an aggregate price increase or an aggregate price decrease for given rates.

[4] In the context of a productivity analysis deflated settled revenues are only approximate as an output measure to the extent the settlement process reflects the various carriers' share of the costs involved in connecting two stations across the territory of two or more carriers.

[5] In many instances this follows from the use the demand models are put to. Typically they are designed as one input toward the construction of an overall production model. Since most authors have utilized their demand model to justify their hypothesis regarding profit maximization on some of the firm's outputs, namely toll services, and since the econometric analysis of production has not been able to tackle more than three outputs - three inputs production or cost functions these authors have been forced to restrict themselves to an aggregate message toll demand curve.

6 The evidence is not as clear in the context of Bell Canada (1980) as that model is far too aggregated to draw this kind of conclusion.

7 Whenever this is not the case, then this form of externality becomes very relevant. Two examples would be the difference in rates in a call between Canada and the USA (Europe) depending whether it originates in Canada or in the USA (Europe) and the GTE USP experiment in Illinois. In the latter, the usage charge of a call originating from a multiparty line is zero, which is not the case for most calls originating from individual lines. That situation does not apply to B.C.-Alberta calls. As long as there is a complete uniformity in rates, then the incentive to shift calls is minimized, even though differences in income, social or demographic characteristics could still be expected to play some role.

8 Even in a system where penetration rate increases very rapidly, one could expect to observe this phenomenon. Practical examples such as France in the last few years would lead one to believe that this is not the case. Usage per main stations has been decreasing in France, the explanation being given by the PTT being that, as the population rate increases, penetration is progressively within lower income groups who either for lack of habit or for income reasons use the network much less.

9 It is contended here that the Canadian market is fundamentally saturated, yet it is recognized that saturation is a vague concept. Thus Bell Canada gives the number of residential main stations per person 15 and over (B-81-206), a series which still exhibits growth, even if it is at a slightly decreasing rate. It also shows the total number of business telephones per person employed (B-81-212) which also exhibits continued growth. Both observations, however, while

useful from a marketing point of view, cannot be used to reject the hypothesis that saturation generally characterizes the Canadian market. Hence a more appropriate measure, for the residence market, would be to express main stations in terms of households, the size of which has been changing with the age composition, etc. Similarly, if in the context of business services, one considers the trend toward service industry, and if one excludes new services such as data transmission,...it is likely that, once again, there is near saturation in any one industry.

[10] For a counter argument, see Curien and Vilmin (1981)

[11] For definitions and description of variables, see Table A.3 in Appendix.

[12] This is distinct from assuming that calls should be uniformly distributed, independently of distance. The latter is not assumed in this paper.

[13] We owe this point to Jon Breslaw.

[14] This condition may change as the network is put to new usage such as data transmission etc.

[15] There are other definitions of price such as a chained price index that can be employed. A chained price index may be obtained from repricing a given volume with new rates each time they come into effect. Unfortunately, such a price index is not available at present.

[16] Two attempts were made in the estimation to account for this problem. First, for the quarters that were affected, the average prices for the other two months were used instead, and second a dummy variable was inserted into the equation. Unfortunately, the results are not as expected with some of the coefficients having the wrong signs and are generally less significant than those reported later on in Tables 8, 9 and 11.

[17] Two alternatives have been considered: a variance components model and a first-order autoregressive model with contemporaneous correlation. These models are described in Drummond and Gallant

(1979) who implemented the estimation procedure for SAS. Attempts to estimate the first model were unsuccessful because of insufficient cross-sectional observations. Attempts to estimate the second model produced results quite similar to that of OLS. These results are reported in Tables A.4 and A.5. Briefly, all the price and income coefficients have the expected signs but the coefficients are now less significant. There is however one significant difference; the income squared coefficient is now much smaller in magnitude.

19 In fact for revenue forecasting following a proposed rate change, it would become necessary to forecast or extrapolate the elasticities.

18 Because the translog and the double log model are not nested, an attempt was also made to test for the truth of each model with the other as the alternative hypothesis, the Davidson-MacKinnon (1981) J test for non-nesting hypothesis was applied. The double log model used is equation (15) rewritten in its "stack form" so that the number of observations is the same in each model. Unfortunately, the results reject both hypotheses.

20 This is probably due to data problems such as personal disposable income for B.C. which is dominated by incomes in Vancouver and Victoria and is not representative of income of callers in mileage band A.

BIBLIOGRAPHY

(1) Abrahan, Bovas and George E.P. Box (1975): "Linear Models, Time Series and Outlines 3: Stochastic Difference Equation Models", Department of Statistics Technical Report No. 430, University of Wisconsin, Madison.

(2) Bell Canada P(CRTC)23 Dec. 76-500, Tab. K.

(3) Bell Canada (1969): "Productivity Measures" Exhibit No. B242, CTC.

(4) __________ (1973): "Memorandum on Bell Canada Productivity", Exhibit No. B-73-62.

(5) __________ (1980): "Econometric Models of Demand for Selected Bell Canada Services", Attachment 1 in Bell (CRTC)03 Apr. 80-809.

(6) B.C. Tel (1980): "Some comments on Jeffrey I. Bernstein's A Corporate Econometric Model of British Columbia Telephone Company", Mimeo, B.C. Telephone Company, Burnaby.

(7) Bernstein, Jeffrey I. et al (1977): "A Study of the Productive Factors and Financial Characteristics of Telephone Carriers", DGCE Working Paper 52, Department of Communications, Ottawa.

(8) Bernstein, Jeffrey I. (1980): "A Corporate Econometric Model of the British Columbia Telephone Company", Public Utilities Forecasting, ed. O. Anderson, North Holland, Amsterdam.

(9) Blackorby, C., D. Primont and R. Russell (1978): Duality, Separability and Functional Structure: Theory and Economic Applications, North Holland, New York.

(10) Breslaw, Jon A. (1980): "Simulations of Bell Canada Under Various Rate Scenarios", DGCE Working Paper #161, Department of Communications, Ottawa.

BIBLIOGRAPHY (Continued)

(11) __________ and J.B. Smith (1980): "Efficiency, Equity and Regulation: An Econometric Model of Bell Canada", DGCE Working Paper #145, Department of Communications, Ottawa.

(12) ____________________ (1981): "Efficiency, Equity and Regulation: A Model of Bell Canada", presented at this Conference.

(13) ____________________ (1981): "Efficiency, Equity and Regulation: An Econometric Model of Bell Canada", Working Paper #81-01, Department of Economics and Institute of Applied Economic Research, Concordia University, Montreal.

(14) Christensen, L.R., D. Jorgenson and L.J. Lau (1973): "Transcendental Logarithmic Production Frontiers", Review of Economic and Statistics, LV-1, pp. 28 - 45.

(15) Cleveland, William P. (1972): "Analysis and Forecasting of Seasonal Time Series", Ph.D. Dissertation, University of Wisconsin, Madison.

(16) _______________ and George C. Diao (1986): "Decomposition of a Seasonal Time Series: A Model for the Census X-11 Program", Journal of the American Statistical Association. Vol 71, pp. 581-587.

(17) Corbo, V., J.A. Breslaw and J.M. Vrljicak (1978): "A Simulation Model of Bell Canada", DGCE Working Paper #73, Department of Communications, Ottawa.

(18) Corbo, V., J.A. Breslaw, J.M. Dufour and J.M. Vrljicak (1979): "A Simulation Model of Bell Canada: Phase II," DGCE Working Paper, Department of Communications, Ottawa.

BIBLIOGRAPHY (Continued)

(19) Courchesne, Camille, Alain de Fontenay and Jacques Poirier (1981): "An Empirical Study of Seasonality in Econometric Modelling", in Time Series Analysis. Anderson,O.D. and M.R. Perryman, eds., North Holland, Amsterdam.

(20) Diewert, W. Erwin (1971): "Functional Forms for Profit and Transformation Functions", Journal of Economic Theory, 6, pp. 284 - 316.

(21) Dreessen, Erwin A.J. (1977): "The Demand for Intra-B.C. Toll Calling - A Preliminary Report", Costs, Prices and Economics' Working Paper, B.C. Telephone Company, Burnaby.

(22) ________________ (1978): "Elasticity is .....", Costs, Prices and Economics' Working Paper, B.C. Telephone Company, Burnaby.

(25) Fuss, Melvyn and L. Waverman (1978): "Multi-product Multi-input Cost Functions for a Regulated Utility: The Case of Telecommunications in Canada", paper presented at the NBER Conference on Public Regulation, Washington, D.C., December, 1977.

(26) (1981): "The Regulation of Telecommunications in Canada", forthcoming report to the Economic Council of Canada.

(27) Granger, Clive W.J. and Paul Newbold (1977): Forecasting Economic Time Series, Academic Press, New York.

(28) Jorgenson,D.W. and K. Nishimizn (1978): "U.S. and Japanese Economic Growth, 1952-74: An International Comparison", Economic Journal, vol. 88, pp. 707-726.

BIBLIOGRAPHY (Continued)

(29) Kmenta, J. (1971): Elements of Econometrics, New York, MacMillan.

(30) Maddala, G.S. (1977): Econometrics, New York, McGraw-Hill.

(31) Piekaar, Ed (1980): "The Intra B.C. Demand Model: Further Developments", mimeo, B.C. Telephone Company, Burnaby.

(32) Rea, John D., and G.M. Lage (1978): "Estimates of Demand Elasticities for International Telecommunications Services", Journal of Industrial Economics, Vol. XXVI, No. 4, pp. 363 - 381.

(33) SAS Institute (1979): SAS User's Guide: 1979 Edition, SAS Institute, Raleigh, North CArolina.

(34) Taylor, Lester D. (1980): "Telecommunications Demand: A Survey and Critique, Ballinger.

APPENDIX

Table A.1

INTRA BC Model = Long-Run Price Elasticities

				Mileage Band						
				A	B	C	D	F	G	H
Segment										
#01	RES	Mon-Fri	DDD Day	- .16	- .36	- .54	- .88	-1.60	-2.02	-2.31
#02	RES	Mon-Thu	DDD Eve	- .11	- .78	- .88	-1.14	-1.34	-1.41	-1.33
#03	RES	Fri	DDD Eve	- .27	- .80	- .43	- .68	-1.33	- .94	- .75
#04	RES	Sat	DDD Day	- .49	- .18	- .28	- .42	- .74	- .83	- .70
#05	RES	Sat	DDD Eve	- .32	- .51	- .27	- .50	-1.33	-1.02	- .67
#06	RES	Sun	DDD D+E	- .39	- .82	- .67	- .91	-1.75	-1.81	-1.45
#07	RES	Mon-Sun	DDD LNI	-1.08	-1.45	-2.08	-2.48	-1.71	-1.77	-1.72
#08	RES	Mon-Fri	SOH Day	- .69	- .39	- .16	- .33	- .81	-1.49	-1.46
#09	RES	Mon-Fri	SOH Eve	- .81	- .44	- .23	- .45	- .78	- .64	- .72
#10	BUS	Mon-Fri	DDD Day	- .19	- .27	- .73	- .97	-1.09	-1.09	-1.93
#11	BUS	Mon-Fri	SOH Day	- .50	- .18	- .24	- .28	+ .18	+ .39	+ .36
#12	BUS	Mon-Fri	P-P Day	-1.49	-1.22	-1.00	-1.05	- .68	- .27	-1.08

Source: Piekaar (1980)

Notes: The mileage for the mileage bands are (0-20, 21-80,81-180,181-290,291-400, 401-500 and 500+ miles) respectively.

Table A.2

INTRA BC Model = Long-Run Income Elasticities

				Mileage Band						
Segment				A	B	C	D	F	G	H
#01	RES	Mon-Fri	DDD Day	1.69	1.09	2.38	2.51	2.16	1.80	.72
#02	RES	Mon-Thu	DDD Eve	2.62	1.30	2.00	1.46	.78	.80	.38
#03	RES	Fri	DDD Eve	3.22	1.16	2.46	1.75	.43	.15	.68
#04	RES	Sat	DDD Day	1.22	1.22	2.11	2.37	1.17	1.03	1.37
#05	RES	Sat	DDD Eve	1.94	1.08	1.83	1.52	.35	- .04	.67
#06	RES	Sun	DDD D+E	3.25	1.03	2.82	2.66	.84	.17	.15
#07	RES	Mon-Sun	DDD LNI	2.84	.87	2.15	1.15	1.22	.65	- .19
#08	RES	Mon-Fri	SOH Day	1.72	2.48	2.74	2.95	2.60	1.53	2.75
#09	RES	Mon-Fri	SOH Eve	1.16	1.60	1.73	1.98	1.06	1.28	2.81
#10	BUS	Mon-Fri	DDD Day	1.82	.28	1.99	2.47	1.69	1.44	.69
#11	BUS	Mon-Fri	SOH Day	- .22	-1.19	.35	.63	- .08	.90	.71
#12	BUS	Mon-Fri	P-P Day	7.59	2.49	2.93	2.94	.95	.68	3.82

Source: Piekaar (1980)

Notes: The mileage for the mileage bands are (0-20, 21-80,81-180,181-290,291-400, 401-500 and 500+ miles) respectively.

Table A.3

Definition and Description of Variables

I Quarterly current dollar personal disposable income in B.C. Estimated as three month sums of monthly series. Monthly figures are obtained by applying monthly wages and salaries in B.C. series from CANSIM to annual income figures. Annual current dollar income figures are obtained from the B.C. Economics Accounts as the differences between total personal expenditure and current transfers to government.

CPI Vancouver consumer price index (1971=1.00). Quarterly data are calculated as 3 month averages of the monthly series obtained from CANSIM.

N Quarterly population of B.C. Estimated as monthly averages. Monthly figures are derived by applying monthly pattern of population over age 15 for men and women to quarterly population of B.C. All series are obtained from CANSIM and all figures are in thousands of persons.

T Numbers of residence main stations in service at B.C. Telephone. Three month averages of monthly data.

Q B.C. to Alberta Monday to Friday day-time DDD call minutes. Figures are in thousands and are obtained from toll sample data.

P Revenue per call minute. Revenue figures are obtained from toll sample data.

$q = Q/T$

$p = P/CPI$

$Y = I/CPI$

$y = Y/N$

Table A.3 (Continued)

OKTS* = OK Tel Strike dummy (1973Q3-1974Q1)

BCTS* = B.C. Tel Strike dummy (1977Q4-1978Q1)

PS* = Postal Strike dummy (1975Q4 and 1978Q4)

S1-S3 = Seasonal dummies.

*All strike dummies are calculated as the ratio of the number of week-days affected by the strike to the total number of week-days in the quarter.

Table A.4

Trans Log Model

	B Values	T for H:B=0	PROB |+| > 0
CONSTANT	2.02	7.12	0.00
lnp	-1.88	-4.55	0.00
(lnp)	-0.17	-1.71	0.09
(lnp)(ln y)	0.65	1.41	0.16
ln y	2.33	2.49	0.01
(ln y)	-2.11	-0.73	0.46
BCTS	-0.01	-0.07	0.94
OKTS	-0.05	-0.89	0.37
PS	0.44	3.36	0.00
S1	-0.06	-1.93	0.05
S2	-0.09	-2.63	0.01
S3	-0.01	-0.20	0.84
δ_1	-6.03	-26.71	0.00
δ_2	-3.53	-20.63	0.00
δ_3	-1.72	-15.38	0.00
δ_4	-0.96	-19.45	0.00
δ_5	-0.48	-20.57	0.00
δ_6	0.21	22.12	0.00

Degree of freedom for t statistics = 178

PROB |+| > 0 indicates the probability of getting a larger absolute t if B = 0.

Table A.5

Price and Income Elasticities
Trans Log Model: GLS

	A	B	C	D	E	F	G
PRICE ELASTICITIES							
73.Q1	-1.23	-1.45	-1.59	-1.72	-1.76	-1.80	-1.80
MEAN	-1.04	-1.32	-1.47	-1.58	-1.62	-1.65	-1.65
79Q4	-0.87	-1.18	-1.32	-1.42	-1.46	-1.48	-1.48
INCOME ELASTICITIES							
73Q1	1.58	1.99	2.28	2.53	2.60	2.68	2.67
MEAN	0.84	0.79	1.05	1.25	1.32	1.37	1.36
79Q4	-0.19	0.79	1.05	1.25	1.32	1.37	1.36

COMMENTS ON THE THREE PAPERS IN DEMAND ESTIMATION SESSION

M. ISHAQ NADIRI
NEW YORK UNIVERSITY

The three papers in this session address a number of important issues in modelling demand functions for telecommunications; Taylor's paper provides a concise survey of research on this subject; Currieu and Vilman attempt to develop a threshold demand model for access and use of the network; and de Fontenay and Lee update Taylor's survey of the literature and develops a demand function with variable income and price elasticities. I shall note briefly some of the issues in these papers and raise some questions about the analytical methodologies that have been followed.

Taylor's paper which draws heavily on his book on this subject covers a wide set of issues. Taylor distinguishes three special features of demand for telecommunication services. They are: demand for access and use of the network, call and network externalities, and option demand. He points out the poor state of the econometrics of demand modelling, the inadequacies of the data, and the lack of an appropriate theory of dynamics to model the demand for telecommunication services. He reaches the conclusion that based on the empirical results reported in the literature, demand for access is probably highly income and price elastic, local calls are price inelastic but income elastic while toll calls are relatively more price and income elastic.

While I agree with most of Taylor's observations, there are several issues that need further consideration. First, the distinction between demand for access and use seem to be somewhat overdrawn; demand for access is essentially a derived demand for use of telephone service, and access characteristics can be considered as qualitative features to this final demand for telecommunication services. Demand for access and use are basically jointly determined and not a two stage approach as suggested by Taylor. He considers option demand to be an intrinsic characteristic of demand for telecommunication services. But this type of precautionary demand that arises due to uncertainty is a feature of many other assets. The point is that by characterising special features to the telecommunication services we will unduly restrict the application of analytic demand models developed in other fields.

The issue of externalities is critical in the analysis of demand for telecommunication, and further research in this area is certainly needed. Several questions arise that are not specifically addressed in Taylor's paper. One is that if we aggregate over all users the externalities become internalized; another is that new subscribers are likely to take it for granted, ex ante, that almost everyone is a subscriber; also magnitudes of the externalities will differ with different types of services; and finally, introduction of new products makes the concept and measurement of network externalities very problematic. What is needed is to address the question of endogeneity of the size of the network and its rate of expansion.

I concur with Taylor about development of more sophisticated demand models and estimation techniques: it is important to look into the issue of multi tariffs, distinction between duration and initiation of calls, use of better specification and estimation of dynamic factors, and more extensive use of the time series- cross sectional estimation techniques suggest that the possibility of variable price and income elasticities and the potential degree of substitution and complementarity among various telecommunication services should be given serious consideration.

The survey fails to mention the effect of changes in the structure of the economy, the role of technological change and impact of regulation on the level and structure of the demand for telecommunication services. Demand for these services could expand not only because of growth of GNP but also a shift in the structure of the economy toward sectors that are intensive in use of telephones. Technological change affects demand as well, e.g. introduction of DDD expanded considerably the demand for toll. In Taylor's discussion, not enough attention is given to demand for quality of services, nor is there much discussion of the impact of alternative types of communication services on growth of demand for telecommunication services.

Currien and Vilman, in their interesting paper, attempt to determine the threshold income for access and level of use of the network. According to their model, threshold income is determined by three variables: basic tariff, degree of penetrations and fixed tariffs. The network expands as more members of the lower income groups become subscribers and the rate of consumption of telephone services increases with greater penetration. The authors make an important analytical contribution by explicitly introducing income distribution as a determinant of demand for access. Unfortunately, their paper is unnecessarily inaccessible which could be improved by some clarification and condensation of the present paper.

On the substantive issues, Currien and Vilman look at the effect of penetration rate on demand (intensive margin) assuming that income and population remain fairly stable. However in an expanding economy, both these forces will be operative and the net effects of each of them should be separated . Also, the penetration rate itself is essentially endogeneous and its determinants need to be fully specified in the model. Finally, the authors make the strong assumption of separability of the telecommunication services and other goods and services in the consumer's utility function. The consequence and realism of this assumption need to be explored further.

There is a substantial gap between Currien and Vilman's theoretical model and empirical analysis. It is not clear why the particular functional forms such as equations (12) and (33) were chosen for empirical analysis. Nor is there much explanation of the empirical results and the quality of data used in the regression analysis.

de Fontenay and Lee provide a very useful survey of the literature since

Taylor's recent book and argue that most of the researchers have adapted functional forms with constant price used income to estimate on demand for telecommunication services. The authors claim that income and price elasticities for telecommunication services vary with distance and other characteristics and therefore the conventional demand models are misspecified. They formulate a translog demand function which allows these elasticities to vary and plan to use time series-cross section data to estimate it. Their empirical results, unfortunately, are not completed yet and therefore the final judgment on this work has to await future evidence.

COMMENTS ON
TAYLOR, CURIEN AND VILMIN,
AND
DE FONTENAY AND LEE PAPERS

ROBERT A. SPROULE

MANITOBA TELEPHONE SYSTEM

I have three general comments to make here today. The first concerns the Taylor paper. I suggest that Professor Taylor subtitle his paper "A Survey of Results Based On Non-Experimental Data". The point is that Professor Taylor has overlooked recent efforts to estimate demand parameters for experimental data. In particular, Park et al (1980), Park and Wetzel (1981), and Wilkinson (1981) provide examples of the use of experimental methods in measuring the consumer response to changes in the regime used to rate telecommunications services.

If Professor Taylor were to enlarge his survey to include this experimental work, I would like to see the following questions addressed. What factors have given rise to such experimental work? Is the range of estimates of demand obtained from non-experimental data so broad, i.e., the presence of confounding variables so significant, that experimental methods are viewed as the only means of narrowing this range? Conversely, is it the case that structural limitations in the non-experimental rate data are so severe that experiments are required? Finally what are the methodological and financial net benefits of conducting rate experiments for telecommunications services?

My second comment concerns a point which Professor Taylor makes, and a point which Messrs. de Fontenay and Lee may wish to consider. In using the utility maximization paradigm to derive a demand function for message toll service, Taylor (1980, pp. 47 - 50; 1981, pp. 22 - 24) observes that special consideration must be given to the effects of a multipart tariff on the properties of the demand function. In particular, he notes (as do Acton et al. (1980) and Hanoch and Honig (1978) in other areas of applied microeconomics) that a multipart tariff is responsible for the generation of a discontinuous demand function if this function is derived from the utility maximization paradigm. In their rationalization of the choice of a second order approximation to an arbitrary demand function for message toll service, de Fontenay and Lee may wish to mention the above problem, and address it by calling on the argument that Taylor makes: namely, discontinuities in the individual's demand function disappears with aggregation over individuals as long as income or tastes vary across individuals.

Finally, I am very intrigued by the Curien and Vilmin paper. I initially thought that if there was any weakness in the paper it would show up under close inspection of the microfoundations. After an exchange with Monsieur Curien, all potential issues that I had identified have been resolved. I find the Curien and Vilmin microeconomic model is an exciting departure from the usual neoclassical apparatus.

References

Acton, J. P., B. M. Mitchell, and R. Sohlberg, 1980, "Estimating residential electricity demand under declining-block tariffs: An econometric study using micro-data," Applied Economics, Vol. 12, pp. 145 - 61.

Curien, N., and E. Vilmin, 1981, "Demande et consommation telephoniques: Un model residentiel globat," unpublished manuscript, Direction Generale des Telecommunications, Paris.

De Fontenay, A., and M. Lee, 1981, "B.C./Alberta Long Distance Calling," unpublished manuscript, Department of Communications, Ottawa.

Hanoch, G., and M. Honig, 1978, "The labor supply curve under income maintenance programs," Journal of Public Economics, Vol. 9, pp. 1 - 16.

Park, R. E., B. M. Mitchell, and B. M. Wetzel, 1980, "Demographic effects of local calling under measured vs. flat rate service: Analysis of data from the g.t.e. illinois experiment," in Pacific Telecommunications Conference Proceedings, Pacific Telecommunications Conference '80, Honolulu.

Park, R. E., and B. M. Wetzel, 1981, "Charging for local telephone calls: Pricing elasticity estimates from the g.t.e. illinois experiment," The Rand Corporation, Santa Monica, R-2635-NSF.

Taylor, L. D., 1980, Telecommunications Demand: A Survey and Critique, Ballinger Publishing Co., Cambridge.

Taylor, L. E., 1981, "Problems and issues in modelling telecommunications demand," unpublished manuscript, University of Arizona, Tucson.

Wilkinson, G. F., 1981, "The estimation of usage repression under local measured service: Empirical evidence from the g.t.e. experiment," unpublished manuscript, G.T.E., New York.

REMARKS

to the Demand Estimation session of the Conference "Telecommunications in Canada: Economic Analysis of the Industry".

ERWIN A.J. DREESEEN
B.C. TELEPHONE

1. I am grateful to the organizers of this Conference for the opportunity to compensate somewhat for Mr. Lee's forced absence. My remarks first address several points Professor Taylor raises in his paper. In section 2 I compare results under two toll model specifications which will illustrate the reasons for my unhappiness with one of Professor Taylor's recommendations. Some highlights using B.C. Telephone's most recent model for intra-province toll calling follow in section 3. I conclude with some suggestions for research in telephone demand analysis, in addition to those enumerated by Professor Taylor.

Professor Taylor correctly typifies the demand for telephone service as consisting of demand for access and demand for usage. He also gives a useful account of how positive network and call externalities, and the phenomenon of option demand, lead one to expect an equilibrium network size larger than one would observe in their absence. His discussion of the opportunity cost of time points to the advisability of distinguishing volumes of calls and their average duration, even under a uniform-price-per-minute regime.

In his discussion of empirical studies Professor Taylor dismisses those which use as the price variable a ratio of revenues over quantities because such a procedure would necessarily establish a negative relationship between price and volume. This is a potentially very misleading statement, since it is only valid when revenues are completely unrelated to changes in quantity! More generally, Professor Taylor criticizes any ex-post procedure for determining price. But, strictly speaking, almost any price measurement is ex-post since in all but extreme cases a tariff schedule cannot be used directly. It is true, of course, that under a price index procedure such as Laspeyres the relationship between the dependent and the independent variable is attenuated because only base period quantities come into play on the right-hand side.

It is also true that there is indeed a potential for bias, namely if the observed volumes are subject to measurement error. But the existence and direction of that bias then depends on the model form (e.g. whether it is linear or log-linear), on the type of error (e.g. whether it is multiplicative or additive), on whether the same measurement error is present in the revenues data or not, and on whether the true values are above or below the observed values.

In connection with pricing I must point out a minor lapse in Professor Taylor's discussion: Not all of Canada has a multi-part tariff for toll calls. B.C., for one, has had intra-province DDD calling at a uniform-price-per-minute rate for 5 years starting in 1976. Such a regime is also in effect in the territories of Saskatchewan Tel, Maritime Tel, Island Tel and Newfoundland Tel.

In reviewing the empirical literature Professor Taylor appears to accept inclusion of system size among the explanatory variables as an adequate method for gaining insight in the degree of call and network externality. In his monograph he expressed more doubt in this regard when he wrote: "I want to warn readers not to take the models that have been presented too literally. Telephone demand is a complex subject, and there is much that is not known. ... I have tended to roll both the [call and access] externalities and option demand into the size of the system (as measured by the number of subscribers), but this can hardly be considered a satisfactory solution, since it leaves the individual contributions under-identified." [Taylor, 1980, pp. 66-67.] In my view even this cautious attitude does not go far enough: The researcher's expectation regarding externalities (or anything else, for that matter) should be hemmed in by his prior knowledge of the subject of his analysis. Professor Taylor himself, in this paper, notes that endogenous system growth driven by network externality is probably a thing of the past in North America. In any case, a test for the existence of externalities (and option demand) belongs in the demand for access side of a demand for telephone service model: Nothing in the theory presented by Professor Taylor suggests that usage per main station would be affected by externalities or option demand.

This is not to say that a number of hypotheses regarding usage and changing subscriber membership could not be tested. Mr. Curien's paper, for instance, contains the suggestion that newer subscribers are drawn from progressively lower income groups and, under a pay-for-use regime, make less use of their telephone. Or, in an affluent society a reasonable hypothesis could be that new young households indulge more readily in the "long distance feeling". But the important point here is that such hypotheses are obviously not adequately tested by regressing usage volumes on (*inter alia*) the number of main stations. A more sophisticated approach is required. Apart from these refinements a basic requirement for a usage model is that it be made conditional on the number of stations in the system. This is fulfilled in a straightforward manner by adopting the volumes-per-station approach. (1)

(1) In the toll usage-per-station framework one could argue that the divisor should perhaps not be the traditional count of main (and other) stations, but a count of subscribers participating in non-local usage. The toll usage participation rate could be made endogenous by explaining it as a function of income, household composition, EAS growth, toll prices, etc.

Failure to specify some kind of "per capita" model for toll demand may lead to econometric nonsense. An example is documented in the next section. I use monthly intra-B.C. data from 1973 to 1978 for Residential Monday-through-Thursday evening direct-dial calling in the 51-100 mileage band.

Two Models

2. The data are seasonally adjusted by the Census II method, the estimation technique is OLS, the dynamic specification is Koyck distributed lag. Explanatory variables common to both models are a shift variable for a discount-time definition change in February 1975 (EDCH), a dummy variable for the 1975 postal strike (POSTAL), and a dummy variable for telephone workers' strikes in 1973 and 1977-78 (STRIKES). Also common to both models is LSCAL, the log of a standardized and normalized count of Mondays through Thursdays in each month. All following variables are also in logarithms.

Model 1 has conversation minutes (LM) as the dependent variable. Apart from the lagged dependent variable (LM(-1)), the model postulates Revenue per Minute in current dollars (LNP), Personal Disposable Income in B.C. (LNPDI), Vancouver's Consumer Price Index (LVPI), a Residential Main Stations count (LSTA) and B.C. Tel's advertising expenditures in current dollars (LNADV).

Model 2, in contrast, has conversation minutes per Residential Main Station as the dependent variable (LMS), divides Personal Disposable Income by the population of B.C. of age 15 or over, and divides all dollar amounts by Vancouver's Consumer Price Index (LRPDIPC, LRP, LRADV). The respective estimates are exhibited in Table 1.

The overall significance, fit and serial correlation statistics for model 1 are all impressive. The coefficients for LVPI and LNP are close (except for opposite sign), suggesting that customers do not suffer greatly from money illusion. But the equation suffers under three major defects. One, the "habit" coefficient, 0.06, is implausibly low, particularly in the context of a monthly model [(2)].

Table 1

Two models for Weekday Residential Evening DDD Calling, 51-100 Miles, Intra-B.C., Jan. 1973 - Dec. 1978 (N = 72).

Model 1: Dep. Var. = LM

	Constant	LM(-1)	LNP	LNPDI	LVPI	LSTA
b:	-26.8	0.06	-0.64	-0.12	0.48	3.15
$\lvert t \rvert$:	(2.0)	(0.6)	(4.9)	(0.5)	(0.9)	(3.0)

	LNADV	STRIKES	POSTAL	LSCAL	EDCH
b:	-0.00	0.03	0.06	0.82	-0.08
$\lvert t \rvert$:	(0.4)	(1.1)	(2.2)	(8.7)	(3.9)

F(10, 61) = 413 $\bar{R}^2$ = .983   SER = .0372   Durbin's h = -.20

Mean Lag	Long-Run Elasticity of		
(in months)	Price	Income	Stations
0.0	-0.68	-0.13	3.35

Model 2: Dep. Var. = LMS

	Constant	LMS(-1)	LRP	LRPDIPC
b:	1.04	0.44	-0.73	0.62
$\lvert t \rvert$:	(1.6)	(5.1)	(4.7)	(2.4)

	LRADV	STRIKES	POSTAL	LSCAL	EDCH
b:	0.01	0.04	0.03	0.97	-0.11
$\lvert t \rvert$:	(0.9)	(1.4)	(0.8)	(8.6)	(4.1)

F(8, 63) = 175 $\bar{R}^2$ = .951   SER = .0461   Durbin's h = -1.34

Mean Lag	Long-Run Elasticity of	
(in months)	Price	Income
0.8	-1.30	1.11

Two, the coefficient for income obtains the wrong sign. Both these coefficients, as well as the one for LVPI, have low levels of significance. Three, the result for LSTA says that every 1% increase in Stations implies a 3.3% increase in evening conversation minutes. Some network externality! Such a grossly implausible result (if causality is imputed) suggests a basic misspecification of the model.

Not much changes to the estimate of the price coefficient between models 1 and 2. But in model 2 the lagged dependent variable obtains 0.44 and the income elasticity is an ordinary 1.11. The summary statistics, though all less favourable than model 1's, still pass all tests.[3] [4] Under model 2's more plausible dynamic estimate the long-run elasticity of price comes to -1.30, indeed a result with a different policy implication than the -0.68 of model 1.

[2] Professor Taylor has pointed out to me that there is reason to expect a lower "habit" coefficient as the observation period becomes shorter because for some time after a call is made the caller may be satiated and may not call again. This could be true enough for a single call by an individual caller. On a more aggregate level, however, this micro-micro effect is likely to be swamped by observable habit.

[3] Removing LVPI from model 1 and dividing all dollar values in constant terms makes for little change in the estimates or the summary statistics. Only the estimate for LSTA comes down to 2.56, with a t-value of 6.3, and an elasticity of 2.72.

[4] Removal of the "insignificant" variables LRADV, STRIKES and POSTAL from model 2 leads to marginally more significant results, but virtually unchanged b-values. The greatest change is in the elasticity of price, which declines from -1.30 to -1.24.

Recent Disaggregate Evidence from B.C. Tel's Toll Market

3. I believe it is insufficiently realized that aggregate models of toll usage give very little guidance to the product manager who has to decide on how to fulfill a given additional revenue requirement. The Intra-B.C. Toll Demand Model attempts to meet that information need. I'm happy to report that, from the start, we have been telling a consistent story, particularly with respect to price elasticities.(5) This work, incidentally, belies Professor Taylor's contention that the empirical record is at this stage unable to support a distinction between residence and business customers: It can do that, and much more.

(5) Dreessen (1977) found, using impure July '71 - May '76 data in a Hildreth-Lu dynamic specification, both residence and business sectors to have near-zero price elasticities in the 0-30 mileage band. Dreessen (1978, e.g. model Log-Lin I, p. 27) estimated, over July '71 - September '76, the aggregate Residence market to have a price elasticity of -1.26, the Business market -0.82; for the Total market the estimate was -.99. These results agree well with the ranges reported in Dreessen (1979) and Piekaar (1980) for detailed market segments.

Price elasticities for B.C. Tel's four most important markets over 6 mileage bands are exhibited in Table 2. In 1978, these markets represented 55% of intra-B.C. toll volume (6).

(6) Piekaar (1980) estimates a total of 12 market segments over 7 mileage bands, bringing coverage up to 79% in terms of 1978 toll conversation minutes. The estimates in Table 2 are derived with the same observation period, dynamic and variable specification and pattern of grouping (using Zellner's SUR) as in Piekaar (1980), but employed an updated raw data base. The slight differences in estimates are due to the effect of the deseasonalization procedure.

Table 2

Long-run Price Elasticities for 4 x 6 Intra-B.C. Market Segments, January 1973 - December 1978 (N = 72), Logarithmic Koyck Models.

Segment		Percent of 1978 Intra-B.C. Toll Volume	Mileage Bands					
			11-25	26-50	51-100	101-175	176-300	301-870
1	Res. Mon.-Fri., DDD, Daytime	12.0	-.38 (±.25)	-.52 (±.17)	-.88 (±.24)	-1.58 (±.41)	-1.99 (±.55)	-2.28 (±.55)
2	Res. Mon.-Thu., DDD, Evening	15.7	-.78 (±.35)	-.89 (±.32)	-1.14 (±.36)	-1.36 (±.47)	-1.43 (±.55)	-1.27 (±.59)
3	Res. Sunday, DDD, Day & Evening	7.0	-.82 (±.35)	-.67 (±.41)	-.91 (±.53)	-1.73 (±.52)	-1.84 (±.51)	-1.43 (±.49)
4	Bus. Mon.-Fri., DDD, Daytime	20.3	-.33 (±.30)	-.75 (±.30)	-.98 (±.35)	-1.12 (±.41)	-1.15 (±.46)	-1.95 (±.58)

Note 1: Bracketed figures are 1/2 of the 95% confidence interval around the estimates.

Note 2: See Piekaar (1980) for specification and estimation techniques.

It would appear that these estimates follow some expected patterns: They generally increase in absolute value with increasing distance (i.e., expense of the call); above 100 miles they are higher for Residence than for Business calling and higher in full-price than in discounted periods.

It is worth noting that these results do not unambiguously support the contention that the toll market as a whole, or even the total Residential market is price-elastic. I agree with Professor Taylor's educated guess that aggregate toll elasticity is likely to be between 0.5 and 1.0, though in the case of B.C.'s intra-province market the figure is likely to be closer to the latter than to the former.

Modelling is an art, and one must live with restrictions imposed by data, including interdependence of explanatory variables. Price, the Economy and the lagged dependent variable are correlated to some extent in some of these equations. But, as I have indicated, through many changes regarding the data employed, the dynamic form, the choice of variables and the level of aggregation the evidence on price elasticities has proved remarkably robust. (7)

(7) All B.C. Tel's studies have in common the minutes-per-station specification, the method of counting stations and the constant dollar terms of all monetary variables. Dreessen (1979) includes comparisons with some alternative specifications.

Concluding Comments

4. Professor Taylor, in this paper and in his monograph [pp. 174-180], has provided a useful list of problem areas and of priorities in research on telephone demand. I would merely like to add the following to his <u>parnagujuk</u>:

- Disaggregate and aggregate models need to be tied together more closely, perhaps in a Rotterdam-type model. Aggregate data should be consistent indices rather than gross sums and averages.

- The behavioural implications of a specification in terms of minute-miles ought to be explored; if a suitable form could be found then mileage disaggregation in estimation would become unnecessary.

- If nothing else, then the estimates presented in section 3 clearly make the conventional assumption of constant elasticity assailable in aggregate models. A more flexible form, perhaps also allowing price-income interaction, is strongly indicated.

- On long-distance participation rates, at least a comparative static analysis is in order -- this is virtually untrodden terrain.

References

1. Dreessen, E.A.J. (1977), "The Demand for Intra-B.C. Toll Calling", B.C. Telephone Co. (July), 83 pp + Appx., B.C. Tel Rate Case, C.R.T.C. Response to B.C. Tel (BCG) 80 17 18 - 2007A.

2. Dreessen, E.A.J. (1978), "Aggregate Demand for L-D Intra-Province Calling: Estimates and Forecasts, 1971-1977", B.C. Telephone Co. (June), 32 pp + Appx., B.C. Tel Rate Case, C.R.T.C., Response to B.C. Tel (BCG) 80 17 18 - 2007A.

3. Dreessen, E.A.J. (1979), "Elasticity Is ... A disaggregate analysis of intra-B.C. toll calling, 1973-1978", B.C. Telephone Co. (September 18), 18 pp. + Appx., B.C. Tel Rate Case, C.R.T.C. Response to B.C. Tel (BCG) 80 17 18 - 2007A.

4. Piekaar, E. (1980), "The Intra-B.C. Model: Further Developments", B.C. Telephone Co. (January), 14 pp. + Appx., B.C. Tel Rate Case, C.R.T.C., Response to B.C. Tel (BCG) 80 17 18 - 2007A.

5. Taylor, L. (1980), Telecommunications Demand: A Survey and Critique, Ballinger Publishing Company, Cambridge, Mass., 208 pp.