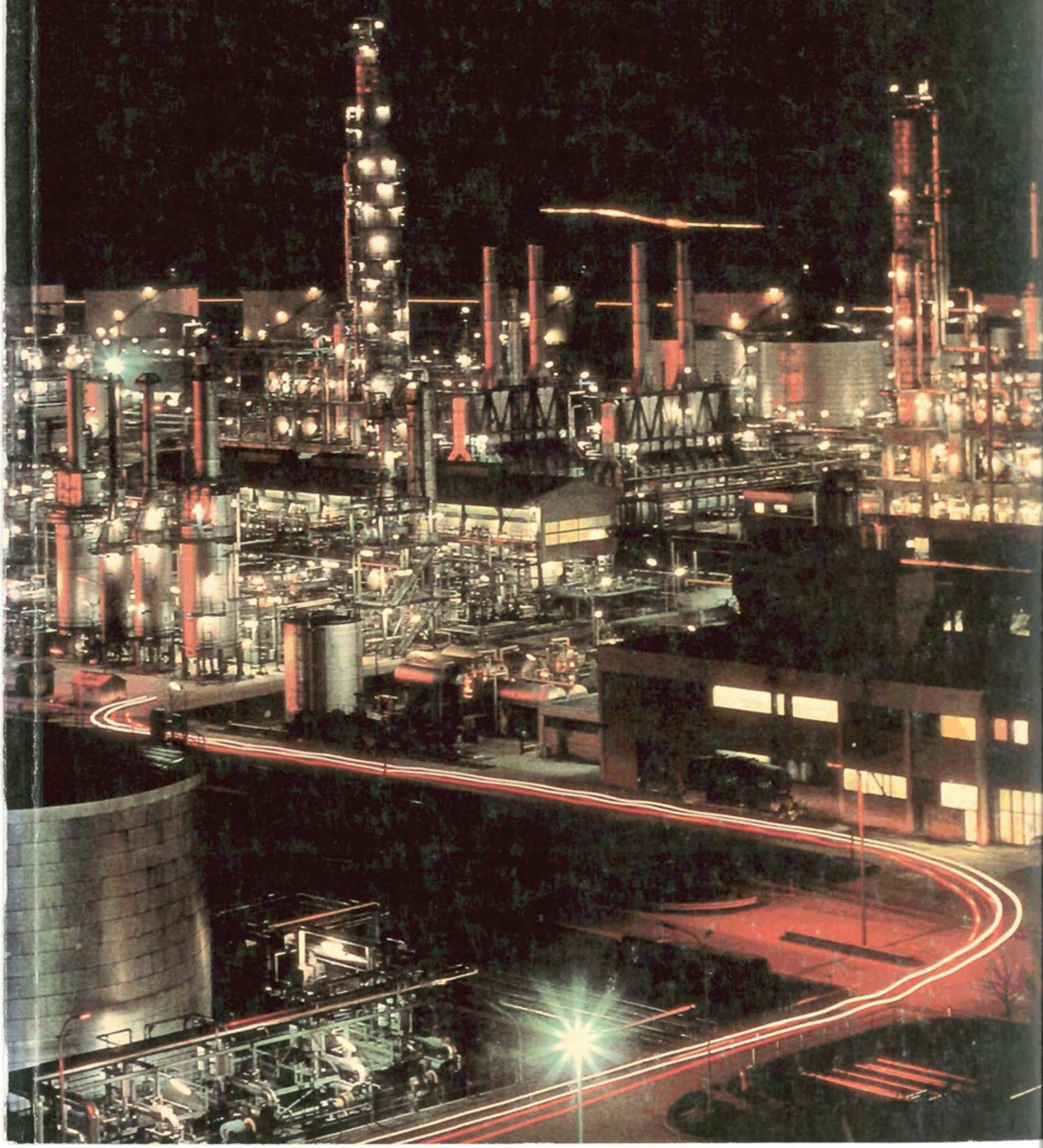
IIASA

REPORTS

International Institute for Applied Systems Analysis

Volume 1 Number 2 April-June 1980
Completing Volume 1



DRAM: A MODEL OF HEALTH CARE RESOURCE ALLOCATION

David Hughes
Andrzej Wierzbicki

PREFACE

The principal aim of health care research at IIASA has been to develop a family of submodels of national health care systems for use by health service planners. The modeling work is proceeding along the lines proposed in the Institute's current Research Plan. It involves the construction of linked submodels dealing with population, disease prevalence, resource need, resource allocation, and resource supply.

This is the second research report on the *d*isaggregated *r*esource *a*llocation sub-*m*odel called DRAM. It describes the extension of the Mark 1 version (RR-78-8) to include the distribution of many resources across different modes of care. The earlier assumption that all available resources must be used has been relaxed, and an extensive analytic treatment suggests various methods for estimating the submodel's parameters. Several case studies that use the model are in progress and reports on these applications will be forthcoming.

This paper is an output of a collaboration between two Areas at IIASA. It describes how a health resource allocation model, developed in the Health Care Systems Task of the Human Settlements and Services Area, may be solved by using optimization techniques studied in the Optimization Task of the System and Decision Sciences Area.

1 INTRODUCTION

It has been widely observed (Feldstein 1967, Van der Gaag *et al.* 1975, Rousseau 1977) that the demand for health care seems to be insatiable. When more hospitals are opened, more patients are treated, and the hope expressed at the inception of the U.K. National Health Service that increasing supplies of health care would reduce subsequent demands has not been realized there or anywhere. The causes of this phenomenon are various, but it gives rise to the same question in all countries: What health care resources should be made available?

Unfortunately, the principal output of health care systems – health – is almost impossible to define or measure (Cardus and Thrall 1977). Much as we would like to design a health care system that would maximize health, we do not even know how to begin. Instead, we seek to predict how those hospitals and other resources available in the health care system (HCS) will be used. Who gets what?

DRAM (a *d*isaggregated *r*esource *a*llocation *m*odel) is designed to help answer such questions. It is one of the submodels of the HCS model conceived by Venedictov *et al.* (1977), and being developed by a group of scientists from different countries working at the International Institute for Applied Systems Analysis (IIASA). Figure 1 shows the five groups of submodels of the HCS developed so far at IIASA; they are explained in more detail in a recent status report (Shigan *et al.* 1979). This figure represents one part of the complete HCS: the processes by which people fall ill and by which resources are provided and used for their care. DRAM (in the group of resource allocation submodels) represents how the HCS allocates limited supplies of resources among competing demands of morbidity. Specifically, it asks *If a certain mix of health care resources (e.g., hospital beds, nursing care) is available, how will the HCS distribute them among patients?* DRAM does *not* prescribe an optimal allocation of resources.

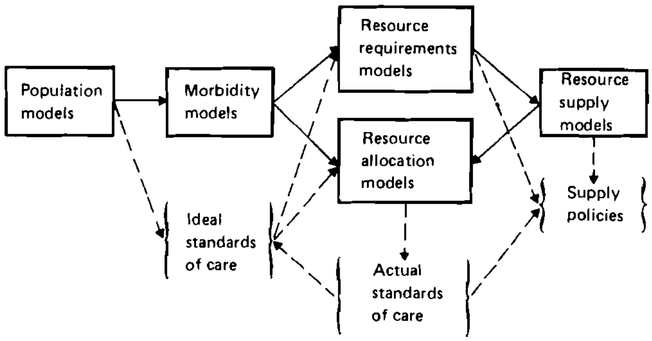


FIGURE 1 The family of HCS submodels constructed at IIASA.

Instead, it *simulates* how the HCS responds when resource availability changes. Even in countries with market economies, there are invariably some planning instruments for controlling the supply of public goods. But even in countries with planned economies, resources cannot be allocated in a rigid, centralized manner. In every country, doctors have clinical responsibility for their patients, and the pattern of care is determined by many local decisions. McDonald *et al.* (1974), Rousseau (1977), and Burton *et al.* (1978) are among those who have modeled this behavior, and DRAM has close links with the first of these models. Other models for health care resource allocation were reviewed by Gibbs (1977) and Nackel *et al.* (1978).

Like many models, DRAM has *accounting* and *behavioral* components. In the accounting in DRAM, different types of resources are distributed among patients

- In different categories (e.g., age, diagnosis)
- In different modes of care (e.g., inpatient, outpatient)
- With different levels of resources per patient (e.g., length of stay in hospital)

and no more resources are allocated than are available. The resources can be determined by a resource supply/production submodel such as the IIASA submodel described in Shigan *et al.* (1979), or they can be set by the user as a trial policy option.

The behavioral assumption in DRAM is that the HCS behaves as if it were maximizing a preference function that increases with the number of patients treated and the resources received by each. Some of the parameters in this function represent demand inputs, like the *ideal levels* at which patients would be treated and would receive resources if no constraints on resource availability existed. These parameters indicate the true "needs" for health care. Other parameters represent the *elasticities* of the actual levels to changes in resource supply, and the balance between need and supply. The relative *costs* of different resources are other parameters used by DRAM to choose between alternative resource mixes. DRAM does not try to include explicitly every behavioral influence that could be active, but to use parameters that can represent the results of all these influences. Because the parameters have meanings outside the model, they can be estimated by methods that do not involve the assumptions underlying DRAM.

Gibbs (1978b) formulated a pilot Mark 1 version of DRAM. This report is the successor, and summarizes progress up to April 1979. Some but not all of the results have appeared in interim IIASA papers, a list of which appears at the end of this report. Much of this report is about the mathematics of the model, and the examples are concerned with hospital services. Our interests, however, are not so restricted. DRAM is designed to model the concept that the HCS balances the desirabilities of more individuals receiving care against higher average levels of care. Such a model should be applicable in many sectors of health care, and perhaps also in other public sectors.

Readers who are uninterested in mathematical details can skip to Section 5 to read about the use of mathematical models in general and to see examples of DRAM. Two examples are presented: one investigates how hospital beds might be used by the HCS, and the other how the balance between inpatient and outpatient care might change. The other parts of the report develop mathematical results that are needed to support such applications. Section 2 solves the simple DRAM and gives three extensions in which certain restrictions applied to the simple model are removed. Not every resource allocation pattern can be simulated by DRAM, so Section 3 investigates its admissible solutions. This is a way of explaining the implications of DRAM's underlying hypothesis. Section 4 presents methods for calibrating DRAM so that it is appropriate for different questions of policy in different regions. The associated computer programs are not described in this report, but Appendix B provides brief details. Section 6 gives a concise summary of the whole report.

2 MODEL FORMULATION AND SOLUTION

The first step in formulating DRAM is to define variables and to make the key assumptions in the model precise. This is done in Section 2.1, and Section 2.2 analyzes a simple version of DRAM in which all the available resources must be used. Three extensions of the model are analyzed in Section 2.3, and Section 2.4 describes a computational method that can be used to solve all four versions of DRAM.

2.1 *Notation and Assumptions*

We use the indices j = patient category ($j = 1, 2, \ldots, J$), k = mode of care ($k = 1, 2, \ldots, K$), and l = resource type ($l = 1, 2, \ldots, L$) in defining the model variables

x_{jk} = numbers of individuals in patient category j who receive resources in mode of care k (per head of population per year)

y_{jkl} = supply of resource type l received by each individual in patient category j in mode of care k

and in writing $\Sigma_j \Sigma_k x_{jk} y_{jkl}$ as the total resources of type l that are allocated (per head of population per year). DRAM seeks to determine $x_{jk}, y_{jkl}\ \forall j, k, l$, within constraints on total resources, so as to maximize a function

$$U(x, y) = \sum_j \sum_k g_{jk}(x_{jk}) + \sum_j \sum_k \sum_l x_{jk} h_{jkl}(y_{jkl}) \tag{1}$$

where

$$g_{jk}(x) = \frac{\sum_l C_l X_{jk} Y_{jkl}}{\alpha_j} \left[1 - \left(\frac{x}{X_{jk}}\right)^{-\alpha_j}\right] \tag{2}$$

$$h_{jkl}(y) = \frac{C_l Y_{jkl}}{\beta_{jkl}} \left[1 - \left(\frac{y}{Y_{jkl}}\right)^{-\beta_{jkl}}\right] \tag{3}$$

C, X, Y, α, β are model parameters (C denotes $\{C_l, l = 1, 2, \ldots, L\}$ and so on).

The monotonically increasing, concave power functions (2) and (3) follow from general assumptions about aggregate behavior in the HCS. They depict the many agents who control the allocation of health care resources as seeking to attain ideal levels of service (X) and supply (Y), but where the urge to increase the actual levels of service (x) and supply (y) decreases with increasing values of x and y, according to the parameters α and β. The costs of different resources (C) are introduced so that marginal increases in U when ideal levels are achieved ($x = X$, $y = Y$) equal the marginal resource costs. This interpretation is a useful way of introducing meaningful parameters into the model, and Section 4 suggests various ways of estimating X, Y, α, β, C in different applications. For the moment, however, we assume these parameters to be known.

Alternative forms for $U(x, y)$ can be suggested, and some were analyzed by Hughes (1978b). Appendix A presents one of these and shows that minor changes can greatly change both the characteristics of model predictions and the ease of solution. Equations (1)-(3) have convenient analytic properties that make it easy to solve this formulation of the model.

2.2 *The Simple Model*

We seek a solution for x, y that maximizes Eq. (1) subject to the constraints

$$0 \leqslant x_{jk} \leqslant X_{jk} \qquad 0 \leqslant y_{jkl} \leqslant Y_{jkl} \tag{4}$$

In this section, we assume that all available resources of type l, R_l, must be used.

$$F_l(x, y) = R_l - \sum_j \sum_k x_{jk} y_{jkl} = 0 \qquad \forall l \tag{5}$$

With Lagrange multipliers λ_l, $\forall l$, we adjoin the equality constraint, Eq. (5), to the function that is to be maximized, Eq. (1), to give

$$H(x, y, \lambda) = U(x, y) + \sum_l \lambda_l F_l(x, y) \tag{6}$$

When certain convexity and concavity assumptions are satisfied (proved below), the values of x, y, λ that solve the primal problem of $\max_{x,y} \min_\lambda H(x, y, \lambda)$ also solve the dual problem of $\min_\lambda \max_{x,y} H(x, y, \lambda)$. The optimal values $\hat{x}$, $\hat{y}$ are readily found to be

$$\hat{x}_{jk}(\lambda) = X_{jk}\,[\mu_{jk}(\lambda)]^{-1/(\alpha_j+1)} \tag{7}$$

$$\hat{y}_{jkl}(\lambda) = Y_{jkl}(\lambda_l/C_l)^{-1/(\beta_{jkl}+1)} \tag{8}$$

where μ_{jk} is a weighted sum

$$\mu_{jk} = \frac{\sum_l C_l Y_{jkl} \nu_{jkl}}{\sum_l C_l Y_{jkl}} \tag{9}$$

of the terms

$$\nu_{jkl} = [(\beta_{jkl} + 1)(\lambda_l/C_l)^{\beta_{jkl}/(\beta_{jkl}+1)} - 1]/\beta_{jkl} \tag{10}$$

and substituting these values into Eq. (6) yields

$$\begin{aligned} \hat{H}(\lambda) &= H[\hat{x}(\lambda), \hat{y}(\lambda), \lambda] \\ &= \sum_j \sum_k \sum_l \frac{C_l X_{jk} Y_{jkl}}{\alpha_j} \{1 - [\mu_{jk}(\lambda)]^{\alpha_j/(\alpha_j+1)}\} \\ &\quad + \sum_j \sum_k \sum_l \frac{C_l X_{jk} Y_{jkl}}{\beta_{jkl}} \{\mu_{jk}(\lambda)^{-1/(\alpha_j+1)}\} \{1 - (\lambda_l/C_l)^{\beta_{jkl}/(\beta_{jkl}+1)}\} \\ &\quad + \sum_l \lambda_l F_l[\hat{x}(\lambda), \hat{y}(\lambda)] \end{aligned} \tag{11}$$

However, these solutions for x, y are not determined until we find a value $\hat{\lambda}$ that minimizes $\hat{H}(\lambda)$.

In order to see whether this is possible, we inspect the gradient vector of first derivatives $\hat{H}_\lambda$ evaluated at $x = \hat{x}(\lambda)$, $y = \hat{y}(\lambda)$. After much simplification, this is simply the vector with elements

$$\begin{aligned} \frac{\partial \hat{H}(\lambda)}{\partial \lambda_l} &= F_l[\hat{x}(\lambda), \hat{y}(\lambda)] \\ &= R_l - \sum_j \sum_k X_{jk} Y_{jkl} (\lambda_l/C_l)^{-1/(\beta_{jkl}+1)} [\mu_{jk}(\lambda)]^{-1/(\alpha_j+1)} \end{aligned} \tag{12}$$

The corresponding Hessian matrix of second derivatives $\hat{H}_{\lambda\lambda}$ can be written as the sum of two matrices

$$\hat{H}_{\lambda\lambda} = \left\{ \frac{\partial^2 \hat{H}(\lambda)}{\partial\lambda_l \partial\lambda_m} \right\} = A + B \tag{13}$$

with elements

$$a_{lm} = \frac{1}{C_l} \sum_j \sum_k \frac{X_{jk} Y_{jkl}}{\beta_{jkl} + 1} \left(\frac{\lambda_l}{C_l}\right)^{-(\beta_{jkl}+2)/(\beta_{jkl}+1)} [\mu_{jk}(\lambda)]^{-1/(\alpha_j+1)} \delta_{lm}$$

$$b_{lm} = \sum_j \sum_k \frac{X_{jk} Y_{jkl}}{\alpha_j + 1} \left(\frac{\lambda_l}{C_l}\right)^{-1/(\beta_{jkl}+1)} [\mu_{jk}(\lambda)]^{-(\alpha_j+2)/(\alpha_j+1)} \frac{\partial \mu_{jk}(\lambda)}{\partial \lambda_m}$$

where

$$\frac{\partial \mu_{jk}(\lambda)}{\partial \lambda_m} = \frac{Y_{jkm}}{\sum_m C_m Y_{jkm}} \left(\frac{\lambda_m}{C_m}\right)^{-1/(\beta_{jkm}+1)}$$

and where the Kronecker delta δ_{lm} is 1 when l equals m, and 0 otherwise. A is a diagonal matrix with all elements positive. Therefore, any quadratic form $z'Az$ is always positive, as are all the eigenvalues of A. Equivalently, A is positive definite. The matrix B is symmetric, with typical quadratic forms

$$\begin{aligned} z'Bz &= \sum_{lm} b_{lm} z_l z_m \\ &= \sum_{jk} \left[\frac{X_{jk} [\mu_{jk}(\lambda)]^{-(\alpha_j+2)/(\alpha_j+1)}}{(\alpha_j + 1) \sum_m C_m Y_{jkm}} \right] \left[\sum_l z_l Y_{jkl} (\lambda_l / C_l)^{-1/(\beta_{jkl}+1)} \right]^2 \end{aligned}$$

which are non-negative. Therefore, B is positive semidefinite. It follows that $\hat{H}_{\lambda\lambda}$ is symmetric and positive definite, and this guarantees that $\hat{H}(\lambda)$ is strongly convex. Finally, it can be shown that $\hat{H}(\lambda)$ therefore has a unique minimum for some $\lambda = \hat{\lambda}$.

In order to prove that this solution to the dual problem also solves the primal problem, we consider the matrix of second derivatives of $H(x, y, \lambda)$ with respect to the primal variables $z = (x, y)$, evaluated at $x = \hat{x}(\lambda)$, $y = \hat{y}(\lambda)$. In this partitioned matrix

$$H_{zz} = \begin{bmatrix} H_{xx} & H_{xy} \\ H_{yx} & H_{yy} \end{bmatrix}$$

not only the off-diagonal submatrices, but all the off-diagonal terms are zero. The remaining diagonal elements

$$\frac{\partial^2 g_{jk}[\hat{x}_{jk}(\lambda)]}{\partial x_{jk}^2} = -(\alpha_j + 1) \frac{\sum_l C_l Y_{jkl}}{X_{jk}} [\mu_{jk}(\lambda)]^{(\alpha_j+2)/(\alpha_j+1)}$$

$$\frac{\partial^2 h_{jkl}[\hat{y}_{jkl}(\lambda)]}{\partial y_{jkl}^2} = -(\beta_{jkl} + 1) \frac{C_l}{Y_{jkl}} \left(\frac{\lambda_l}{C_l}\right)^{(\beta_{jkl}+2)/(\beta_{jkl}+1)}$$

are negative, so that $H_{zz}[\hat{x}(\lambda), \hat{y}(\lambda), \lambda]$ is negative definite. This is sufficient to ensure that the solution $[\hat{x}(\hat{\lambda}), \hat{y}(\hat{\lambda})]$ is the saddle point to $H(x, y, \lambda)$, and thus solves both dual and primal problems.

It remains to consider the range of possible solutions for λ. As any λ_l tends to zero, all the elements of $\hat{H}_\lambda$ tend to minus infinity. We deduce therefore that $\hat{\lambda}_l > 0$ for all l. In order for the solutions (7) and (8) to satisfy the constraints (4), we should have $\hat{\lambda}_l > C_l$ for all l. Unfortunately, this cannot be guaranteed, and examples can be found that use all the resources but exceed the ideal standards X, Y. These unrealistic solutions are a deficiency of this simple formulation of DRAM which can be overcome by extending the model.

2.3 *Three Extensions*

In the first extension of the simple model, we remove the constraint on individual resource types (5) and add a constraint on total finance. We seek a solution for x, y that maximizes Eq. (1) subject to constraints (4) and

$$F(x,y) = M - \sum_l C_l \sum_j \sum_k x_{jk} y_{jkl} = 0 \tag{14}$$

This solution is the optimal allocation under the assumption that finance M should be used to purchase resources that will maximize the returns of Eq. (1). This assumption is not so realistic for our applications, but it gives a model that is easy to solve.

We find that the optimal values $\hat{x}, \hat{y}$ are the same as solutions (7) and (8), but that the Lagrange multiplier λ is now constant across all resource types, $\lambda_1 = \lambda_2 = \ldots = \lambda_L$. The dual function $\hat{H}(\lambda)$ is a function of a single Lagrange multiplier, λ_1 say, and using the earlier results, we can show that it is the sum of a set of strongly convex functions. It is therefore also strongly convex with a unique minimum for some value $\hat{\lambda}_1 > 0$.

In fact, we can demonstrate a stronger result for this version of the model. Because

$$\lambda_1 = 1 \Longrightarrow \frac{\partial \hat{H}(\lambda)}{\partial \lambda_1} = M - \sum_j \sum_k \sum_l C_l X_{jk} Y_{jkl} < 0$$

$$\lambda_1 \to \infty \Longrightarrow \frac{\partial \hat{H}(\lambda)}{\partial \lambda_1} = M > 0$$

and

$$\frac{\partial^2 \hat{H}(\lambda)}{\partial \lambda_1^2} > 0 \qquad 1 \leqslant \lambda_1 < \infty$$

we deduce that there is a unique optimal value $\hat{\lambda}_1 > 1$ that minimizes $\hat{H}(\lambda)$, provided only that the finance available is less than that required to satisfy all demands $M < \Sigma_j \Sigma_k \Sigma_l C_l X_{jk} Y_{jkl}$. In other words, there is always a unique resource mix that will maximize perceived preferences.

In the second extension of the simple model, we replace the equality resource constraint (5) by an inequality constraint

$$F_l(x,y) - r_l = 0; \qquad r_l \geqslant 0 \qquad \forall l \tag{15}$$

where r_l represents the unused resources of type l, which must always be non-negative. It is easy to show that there always exists a point (x, y, r) that satisfies constraints (4) and (15) provided that the inequality

$$\sum_j \sum_k X_{jk} Y_{jkl} > R_l > 0 \qquad \forall l \tag{16}$$

is satisfied. When sufficient resources of some type are available to violate Eq. (16), it means that there are more than enough of these resources, and that there is no allocation problem! The resource type in excess can be removed from the model.

It is also possible to show that the model can have no solutions with $\hat{x}_{jk} = 0$, $\hat{y}_{jkl} = 0$, or $\hat{x}_{jk} = X_{jk}$. In other words, these constraints are never active. This is because the first two conditions imply that $U(x, y) = -\infty$, and because the last condition requires $\lambda_l = 1$, $\forall l$, which causes constraint (15) to contradict (16). We conclude then that the only constraints that can be active are the upper constraint on y and the lower constraint on r.

There are now just two possibilities. The first possibility is that $\hat{y}_{jkl} < Y_{jkl}$ for all l. Inspection of the function

$$H(x, y, \lambda) = U(x, y) + \sum_l \lambda_l \left(R_l - r_l - \sum_j \sum_k x_{jk} y_{jkl} \right)$$

shows that it is maximized when r_l is zero for all l. The problem is then identical to that analyzed above, and all the previous results hold true. The second possibility is that $\hat{y}_{jkl} = Y_{jkl}$ for one or more (but not L) resource types l. From Eq. (8), the associated values of (λ_l / C_l) are unity, and the rest of the problem is equivalent to the dual problem specified in Section 2.2, but with the extra constraint

$$\lambda_l \geqslant C_l \qquad \forall l \tag{17}$$

The third extension of the simple model subtracts the costs of the used resources from the preference function

$$U(x, y) = \sum_j \sum_k g_{jk}(x_{jk}) + \sum_j \sum_k \sum_l x_{jk} h_{jkl}(y_{jkl}) - \sum_j \sum_k \sum_l C_l x_{jk} y_{jkl} \tag{18}$$

Other things being equal, the model now tries additionally to maximize the value of unused resources. The optimal values of x, y are similar to solutions (7) and (8)

$$x_{jk}(\lambda) = X_{jk} [\mu_{jk} (\lambda + C)]^{-1/(\alpha_j + 1)} \tag{19}$$

$$y_{jkl}(\lambda) = Y_{jkl} \left(\frac{\lambda_l + C_l}{C_l} \right)^{-1/(\beta_{jkl} + 1)} \tag{20}$$

and we may show that $\hat{H}(\lambda)$ is strictly convex as before, with a unique minimum that now lies in the range $\hat{\lambda}_l > -C_l$ for all l. Should we also wish to replace the equality resource constraint by the inequality constraint (15), the appropriate version of the dual constraint (17) becomes

$$\lambda_l \geqslant 0 \qquad \forall l \tag{21}$$

Note that all three extensions of the simplest model have solutions that are transformations of the simplest solution.

2.4 *Solution Procedure*

So far we have demonstrated only that all the versions of the model discussed above can be solved by solving equivalent dual problems. In each case we have to find $\hat{\lambda}$ so as to minimize $\hat{H}(\lambda)$, sometimes subject to constraints like (21), but with a unique solution always guaranteed. Because we know the gradient vector $\hat{H}_\lambda$ and the Hessian matrix $\hat{H}_{\lambda\lambda}$, we can begin to search for $\hat{\lambda}$ by an iterative technique $\lambda^{i+1} = \lambda^i + td^i$ (the upper index i denotes the iteration number) which finds better approximations λ^i, $i = 1, 2, \ldots, N$, to the solution $\hat{\lambda}$, by taking steps with step-size coefficient t, in the Newton direction

$$d^i = -(\hat{H}^i_{\lambda\lambda})^{-1}\hat{H}^i_\lambda \tag{22}$$

Just two refinements are necessary: first, to control the step size, and second, to modify the direction when a constraint like (17) or (21) is applied and encountered.

In order to control the step-size coefficient, we need only reduce it if a step seems likely to overshoot either the solution or a constraint. Figure 2 depicts an appropriate method that tests for this. To proceed when a constraint like (21) is encountered, we determine the set of resource type indices

$$\bar{L} = \left\{ l : \lambda_l = 0, \frac{\partial \hat{H}(\lambda)}{\partial \lambda_l} > 0 \right\}$$

where the constraint is active, and where $\hat{H}(\lambda)$ can decrease only with negative λ_l. The gradient vector $\hat{H}_\lambda$ and the Hessian matrix $\hat{H}_{\lambda\lambda}$ are then projected onto the space of active constraints by replacing all the elements corresponding to active constraints by zeros. They become the *reduced* gradient vector and Hessian matrix and they determine the Newton direction (22) in the space of inactive constraints $l \notin \bar{L}$, which is complemented by zeros for $l \in \bar{L}$.

Figure 3 shows the complete procedure for determining the optimal $\hat{\lambda}$, and hence the solutions $\hat{x}(\hat{\lambda})$, $\hat{y}(\hat{\lambda})$. A matrix inversion is the only potentially difficult computation. Generally, however, the number of different resource types will be sufficiently small (less than five, say) to prevent problems. (Occasionally in the solution of a badly conditioned problem, a step in the Newton direction will not reduce the function $\hat{H}$ because of numerical errors, and steepest descent $d = -\hat{H}_\lambda$ may be necessary.) Note that there is not too much extra computation introduced by an inequality resource constraint. Most

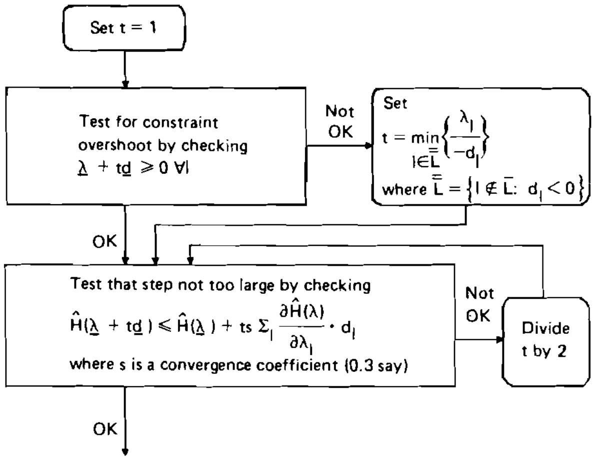


FIGURE 2 Procedure for determining step-size coefficient t.

of the additional refinements are logical rather than computational. All our applications have been solved by a fairly compact computer program, using no special software; Appendix B gives more details about program size and computing efficiency. This program can handle the simple model and all three extensions. In our examples, however, and in most of this report, we refer to the simple model.

3 SOLUTION CHARACTERISTICS

DRAM cannot simulate all patterns of resource allocation that might be observed, and the possibilities for use depend upon the variety of patterns that can be simulated. The analysis given here of admissible solutions to DRAM is restricted to the simplest possible DRAM with one patient category, one treatment mode, and one type of resource, for which all the variants described in Section 2 are identical. Section 3.1 shows how the simplest model can be represented graphically, and gives a fundamental condition on admissible solutions. The results indicate the characteristics of solutions for more complex DRAMs, and suggest ways to fit the model to small numbers of data points. Sections 3.2 and 3.3 derive conditions for fitting two parameters to two data points (Appendix C derives conditions for fitting four parameters to four data points), Section 3.4 derives conditions for fitting two parameters to one data point, and Section 3.5 derives conditions for fitting four parameters to two data points. These results introduce the next section on parameter estimation from many data points.

3.1 *The Simplest DRAM*

For the simplest possible DRAM with $J = K = L = 1$, many elements of the problem can be depicted graphically. First, we can eliminate the Lagrange

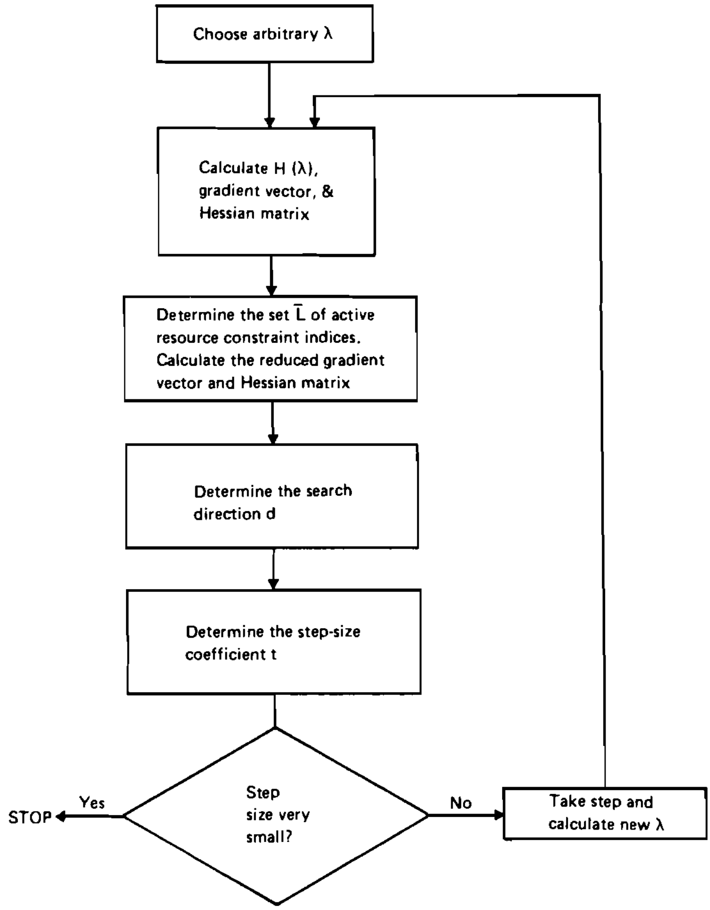


FIGURE 3 Iterative procedure for solving DRAM.

multiplier between Eqs. (19) and (20), to show how the resource level R (which is input to the model) determines the number of individuals treated x, and the supply level y (which are outputs)

$$\left(\frac{R}{XY}\right) = \left(\frac{x}{X}\right)\left[\frac{\beta(x/X)^{-(\alpha+1)} + 1}{\beta + 1}\right]^{-1/\beta} \tag{23}$$

$$\left(\frac{R}{XY}\right) = \left(\frac{y}{Y}\right)\left[\frac{(\beta + 1)(y/Y)^{-\beta} - 1}{\beta}\right]^{-1/(\alpha+1)} \tag{24}$$

It is easy to show that these equations have the shapes shown in Figures 4 and 5. Both curves are concave and monotonically increasing.

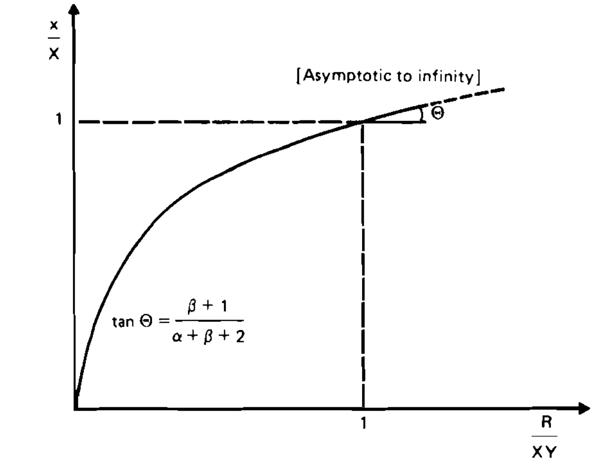


FIGURE 4 (x/X) as a function of (R/XY).

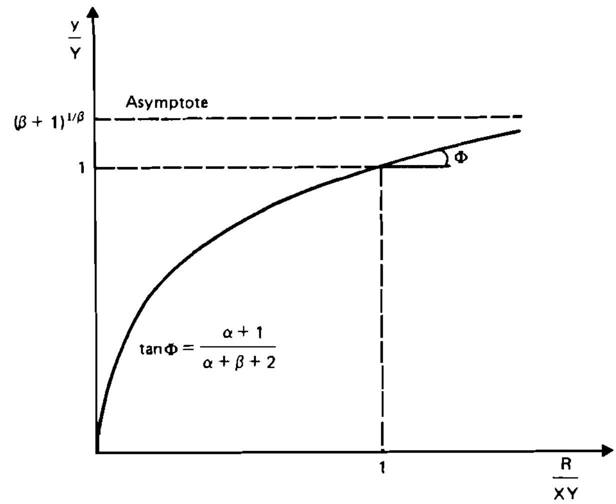


FIGURE 5 (y/Y) as a function of (R/XY).

Alternatively, we can find an equation that relates x and y directly. The result

$$q = \left[\frac{\beta + 1}{\beta p^{-(\alpha+1)} + 1}\right]^{1/\beta} \tag{25}$$

where $p = (x/X)$ and $q = (y/Y)$ is plotted for various ranges of α, β in Figure 6. For $\alpha > \beta - 1$, the curve always has just one point of inflection, and when

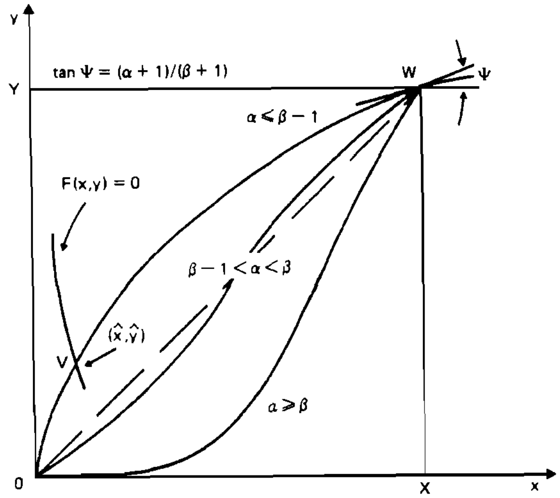


FIGURE 6 Loci of possible solutions on the x–y plane.

$\beta - 1 < \alpha < \beta$, there is just one intersection with the diagonal. From Eq. (25)

$$\frac{dq}{dp} = \left[\frac{(\alpha+1)(\beta+1)}{p^{\alpha+2}(\beta p^{-(\alpha+1)}+1)^{\beta+1}}\right]^{1/\beta} \geqslant 0$$

whence we deduce that two data points (p_1, q_1), (p_2, q_2) can be solutions of DRAM only if

$$p_2 > p_1 \Longleftrightarrow q_2 > q_1 \tag{26}$$

This is a fundamental condition on admissible solutions, which we assume for the rest of this section. It means, for example, that the model cannot reproduce increasing available hospital beds and decreasing lengths of stay, simultaneously. (How this condition should be modified when there are two or more resources, perhaps some increasing and others decreasing, is not clear.)

Equation (25) is the locus of solutions of DRAM on the x–y plane. The particular solution for a given resource level is found at the intersection of the locus and the resource hyperbola $F(x, y) = R - xy = 0$, and it is the point on the hyperbola that maximizes the function of Eq. (18). Figure 7 depicts the shape of this function above the x–y plane. We see that

1. $U(X, Y) = 0$, and $x \to 0$ or $y \to 0$ implies $U(x, y) \to -\infty$. Within the constraints $0 < x < X$, $0 < y < Y$, $U(x, y)$ is always negative and concave.
2. $U(x, Y) = g(x)$ and $U(X, y) = h(y)$. Above the point (X, Y) the surface has gradients

$$\left.\frac{\partial U}{\partial x}\right|_{X,Y} = CY \qquad \left.\frac{\partial U}{\partial y}\right|_{X,Y} = C$$

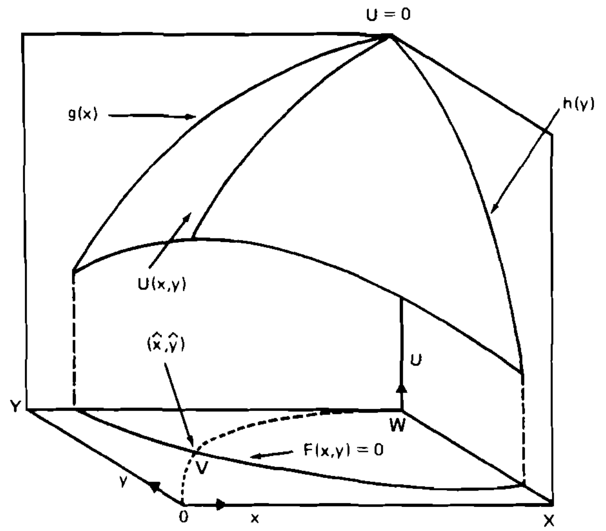


FIGURE 7 The surface $U(x,y)$ above the x–y plane.

3. There is always a unique solution $(\hat{x}, \hat{y})$ because constant-U contours are always more concave than constant-F curves.
4. Equation (25) is represented by the line 0VW.

Evidently, it is not always possible to choose parameters X, Y, α, β that will cause the solution locus 0VW to pass through an arbitrary set of data points (x_i, y_i), $i = 1, 2, \ldots, N$. In the rest of this section we investigate the conditions that allow this.

3.2 *Conditions for Fitting X, Y to Two Data Points*

It seems reasonable that at least two points on Figure 6 are needed to specify a solution locus defined by two parameters, although not all such data will be sufficient or consistent. In this section, we assume that α, β are given, together with two data points (x_1, y_1), (x_2, y_2) such that $x_1 > x_2$, $y_1 > y_2$. Can we choose X, Y such that DRAM can reproduce these points? By substituting the two points into Eq. (25), we easily obtain

$$X = x_1 \beta^{-1/(\alpha+1)} \left[\frac{(y_1/y_2)^\beta - 1}{(x_1/x_2)^{\alpha+1} - (y_1/y_2)^\beta} \right]^{1/(\alpha+1)} \tag{27}$$

$$Y = y_1(\beta + 1)^{-1/\beta} \left[\frac{(x_1/x_2)^{\alpha+1} - 1}{(x_1/x_2)^{\alpha+1} - (y_1/y_2)^\beta} \right]^{1/\beta} \tag{28}$$

The two numerator terms are always positive, and the denominator term is positive if

$$\frac{\beta}{\alpha + 1} < \frac{\ln x_1 - \ln x_2}{\ln y_1 - \ln y_2}$$

This is, therefore, a necessary and sufficient condition for being able to choose X, Y to fit two data points.

3.3 *Conditions for Fitting α, β to Two Data Points*

Alternatively, we can assume that X, Y are given, together with two data points (x_1, y_1), (x_2, y_2) such that $x_1 > x_2, y_1 > y_2$, and ask whether we can choose α, β such that DRAM can reproduce these points. A necessary and sufficient condition for the existence of $\beta > 0$ is easy to find. Writing Eq. (25) as

$$\alpha + 1 = \frac{\zeta(\beta, q)}{-\ln q} \tag{29}$$

where $\zeta(\beta, q) = \ln [(1 + (1/\beta))q^{-\beta} - (1/\beta)]$ we can use the two given data points to eliminate α, giving

$$\zeta(\beta, q_2) - \omega\zeta(\beta, q_1) = 0 \tag{30}$$

where $\omega = (\ln p_2)/(\ln p_1) > 1$. The solution of Eq. (30) is depicted in Figure 8 as the intersection of two curves with known intercepts and asymptotes. There is an intersection for some $\beta > 0$, if $\omega \ln(1 - \ln q_1) > \ln(1 - \ln q_2)$ and $-\omega \ln q_1 < -\ln q_2$, and these conditions can be combined as

$$\tau < \ln y_1 - \ln y_2 < \left(1 + \frac{\tau}{\omega - 1}\right)^{\omega} - \left(1 + \frac{\tau}{\omega - 1}\right) \tag{31}$$

where $\tau = (1 - \omega) \ln q_1$.

A necessary and sufficient condition for the existence of $\alpha > 0$ also comes from Eq. (29). We require that $\zeta(\beta, q_i)/(-\ln p_i) > 1, i = 1, 2$. Unfortunately, it is not easy to remove the dependence on β in this condition. But the two inequalities $\zeta(\beta, q) \geqslant \ln(1 - \ln q)$ and $\zeta(\beta, q) > -\beta \ln q$ lead to two alternative sufficient (but not necessary) conditions

$$q_i < \exp\left(1 - \frac{1}{p_i}\right) \tag{32}$$

$$q_2 < p_2^{1/\beta} \tag{33}$$

where we have used the fact that the second condition is stronger for $i = 2$. We can find a lower bound on β by inspecting the intercepts and asymptotes in Figure 8

$$\ln(1 - \ln q_2) + \beta_{\min}(-\ln q_2) = \omega \ln(1 - \ln q_1)$$

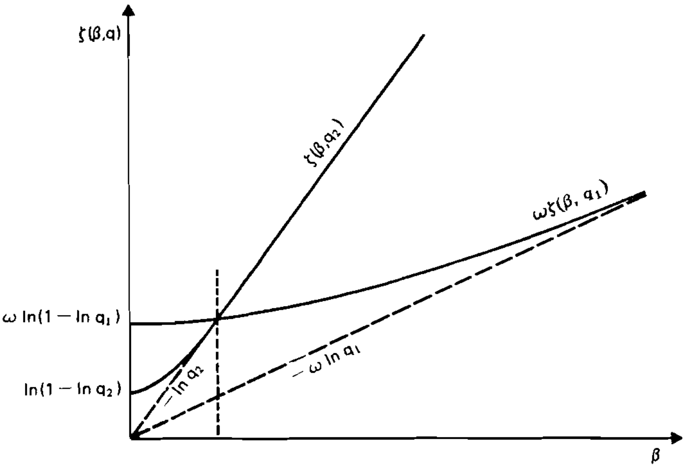


FIGURE 8 Solution of Eq. (30).

When this is used in Eq. (31), the two sufficient conditions become

$$q_i < \exp\left(1 - \frac{1}{p_i}\right)$$

$$\frac{\ln(1 - \ln q_2)}{\ln p_2} - \frac{\ln(1 - \ln q_1)}{\ln p_1} > 1 \tag{34}$$

Empirical evidence suggests that the first condition (32) is less restrictive, at least for small values of (x_1/x_2), and hence closer to being necessary.

These results suggest the following question. Given four points on Figure 6, can we align the solution locus through them all? In other words, given four data points (x_i, y_i), $i = 1, 2, 3, 4$, with $x_1 > x_2 > x_3 > x_4$ and $y_1 > y_2 > y_3 > y_4$, can we choose the four parameters X, Y, α, β such that DRAM can reproduce these points? Sufficient conditions for this, together with an iterative procedure for finding the best fit, are developed in Appendix C. The important conclusion is that even when we have the same number of data points as unknown parameters, and even if the data points satisfy the fundamental condition (26), a perfect fit of the model to the data is not always possible.

3.4 *Conditions for Fitting X, Y or α, β to One Data Point*

In Section 4, we use many data points to estimate pairs of parameters (e.g., X, Y) by combining the estimates suggested by individual data points. We would expect the conditions for fitting two parameters to one data point to be weaker than the conditions derived in Section 3.2 for fitting two parameters to two data points. But is one data point more or less than sufficient to determine two parameters?

In fact, when α, β are given, it is possible to choose an infinite number of pairs X, Y to fit a single data point. Equation (25) shows that for any choice of $X \geqslant x$, there exists some consistent value of $Y \geqslant y$. Similarly, when X, Y are given, it is possible to choose an infinite number of pairs α, β that satisfy Eq. (25) and that therefore fit a single data point. There is, however, a restriction on the minimum possible values of α, β:

$$\alpha_{\min} = \frac{\ln(1 - \ln q)}{-\ln p} - 1, \qquad \zeta(\beta_{\min}, q) = -\ln p$$

which can both be zero, only if $p(1 - \ln q) = 1$.

3.5 *Conditions for Fitting X, Y, α, β to Two Data Points*

Although we do not need the result later, it is interesting to extend and conclude this analysis by asking whether all four parameters can be chosen to fit just two data points (x_i, y_i), $i = 1, 2$; $x_1 > x_2$; $y_1 > y_2$. We analyze this problem in two stages. First, can we choose X, Y so as to satisfy Eq. (31), the necessary condition for the existence of $\beta > 0$? Second, can we also satisfy Eq. (32) or Eq. (34), the sufficient conditions for the existence of $\alpha > 0$?

In order to show that we can always choose X, Y consistent with a $\beta > 0$, we let $\omega \to \infty$ in Eq. (31) giving

$$\tau < \ln(y_1/y_2) < \exp(\tau) - 1 \tag{35}$$

which can always be satisfied by some $\tau > 0$. In practice, ω can be made sufficiently large by setting X close to x_1, and the choice of τ then determines Y.

In order to apply a similar procedure to the sufficient conditions (32) and (34), we write them in the forms

$$\left(\frac{x_1}{x_2}\right)^{\omega/(\omega-1)} - 1 - \frac{\tau}{\omega - 1} < \ln y_1 - \ln y_2$$

$$\ln y_1 - \ln y_2 < \left(1 + \frac{\tau}{\omega - 1}\right)^{\omega} \left(\frac{x_2}{x_1}\right)^{\omega/(\omega-1)} - \left(1 + \frac{\tau}{\omega - 1}\right)$$

where we have set $i = 2$ in Eq. (32). Arguing as earlier that we can choose X to make ω arbitrarily large, we let $\omega \to \infty$ in these equations

$$\left(\frac{x_1}{x_2}\right) - 1 < \ln y_1 - \ln y_2 \tag{36}$$

$$\ln y_1 - \ln y_2 < \left(\frac{x_2}{x_1}\right) \exp(\tau) - 1 \tag{37}$$

Combining Eqs. (35), (36), and (37), we have the sufficient condition

$$\frac{x_1}{x_2} < \min\left\{1 + \ln\left(\frac{y_1}{y_2}\right), \left(\frac{y_1}{y_2}\right)\left[1 + \ln\left(\frac{y_1}{y_2}\right)\right]^{-1}\right\} \tag{38}$$

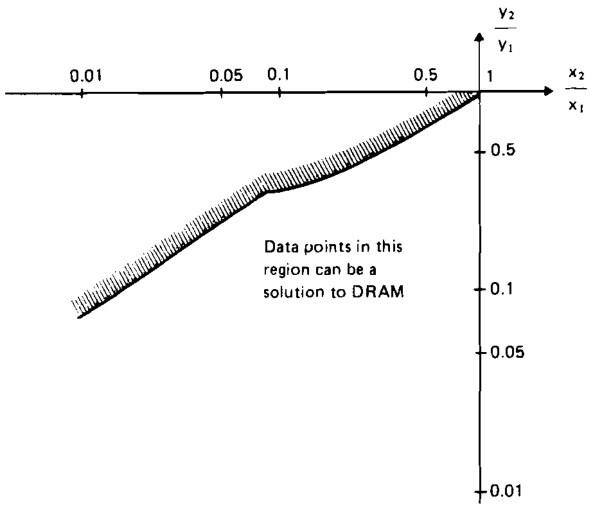


FIGURE 9 Sufficient condition for finding X, Y, α, β consistent with two data points (logarithmic scales).

Figure 9 shows the region of $(x_2/x_1,\ y_2/y_1)$ in which a consistent choice of the four parameters X, Y, α, β is always possible.

This analysis shows that two arbitrary data points, even when they satisfy condition (26), may not be consistent with any choice of DRAM parameters X, Y, α, β. It suggests, therefore, that simple procedures that estimate parameters from just two data points (Hughes 1978a) may be unsuccessful. For this reason, we turn to more general methods for parameter estimation.

4 ESTIMATION OF PARAMETERS

We turn to the problem of calibrating the model, that is, of estimating parameters for DRAM appropriate for a given region and policy question. Section 4.1 reviews sources of data. Sections 4.2 and 4.3 then describe separate procedures for estimating the two pairs of parameters X, Y and α, β, which are drawn together in Section 4.4. These procedures are quite suitable for small examples and they are illustrated in Section 5. Section 4.5 outlines an alternative approach to parameter estimation that incorporates specific assumptions about the uncertainty of model predictions. It shows that, with certain approximations, the approach is feasible and worth testing. Section 4.6 concludes by briefly mentioning the problems of estimating resource costs.

4.1 *Parameters and Data*

The parameters of the model fall into three groups:

- The *ideal levels* X, Y at which patients would be admitted and receive resources, if there were no constraints on resource availability. Absolute values of these parameters have little meaning, but relative values can be chosen to indicate the relative "needs" for health care.
- The *power parameters* α, β which reflect the elasticities of the actual levels to changes in resource supply. For example, we expect the elasticity of admission rate to bed availability to be less for appendicitis patients than for bronchitis patients, because appendicitis usually requires faster attention.
- The *relative costs* C of different resources. DRAM uses the marginal unit cost of a bed-day, a doctor-hour, and so on, or equivalent parameters, in order to choose between alternative mixes of these resources. We defer discussion of resource cost estimation until Section 4.6.

The level of available resources is not regarded as a model parameter but as an experimental variable. DRAM shows how the levels of satisfied demand vary with changes in resource supply.

There are more data available to estimate X, Y, α, β than there are for many other problems in HCS modeling. The sources include:

- Other models
- Special surveys
- Professional opinions
- Routine statistics

At IIASA, *other models* have been developed for other components of the HCS, and particularly for the estimation of true morbidity from degenerative diseases (Kaihara *et al.* 1977) and infectious diseases (Fujimasa *et al.* 1978). Later at IIASA, these outputs may be useful for setting the ideal rates at which patients in different categories need care. Initially, however, we wish to test and use DRAM independently of other models. Many researchers have performed important and useful *special surveys*. Among others, Newhouse and Phelps (1974) and Feldstein (1967) have estimated both elasticities in hospital care and the costs of acute services, and some of these results were used to calibrate a Mark 1 version of DRAM (Gibbs 1978b). Unfortunately, these results may not be relevant in other regions or countries, or at other times. In an international setting it is necessary to avoid relying on results related to a specific health system.

The *professional opinions* of doctors and health planners can be useful for setting ideal levels of care. Countries where there is a high degree of central planning often set normative figures for ideal hospitalization rates and necessary standards of care, and these can be used in DRAM. However, these are not available in all countries, and probably no professional should be asked to estimate elasticities, in case he supplies his own rather than those of the HCS. This leaves *routine statistics*. Most systems keep regular records on the use and costs of their services, and on how they have allocated resources in the past. If DRAM is a valid model of the HCS, then these figures are typical outputs of the model, which we should be able to use for model calibration.

The aim of DRAM is to model how the HCS reacts to change. Generally, therefore, DRAM's model parameters must be estimated from data that show how an unchanging HCS reacts to external changes, either in space or time. *Cross-sectional* data from subregions of the region of interest may show the HCS operating at different resource levels. So also may *longitudinal* data collected at different times. In both cases, however, the underlying system may be different for the different data. Subregions are often deliberately defined so as to be predominately urban or predominately rural, and we must consider ways of averaging the results across the region. Data collected at different times are highly likely to be affected by historic trends in medicine or management. Ideally, we should model these trends and incorporate the time-varying parameters in a time-dependent model. More probably, we shall use data from a period during which we can assume time variations to be small. The resulting model will still be good for representing those aspects of resource allocation behavior that are independent of time trends. A final and obvious problem is that the available data may be incomplete, either because of recording failures or because the data is insufficiently disaggregated.

Not all of these problems can be overcome simultaneously. But in the next three sections we concentrate on estimation methods that are based on routine statistics about current or past allocation behavior, and that take into account that cross-sectional and longitudinal data may reflect inherent parameter variations. In addition, one of the procedures can be used with incomplete data.

4.2 *Estimation of X, Y*

We consider first the estimation of the ideal service levels X and the ideal supply levels Y, assuming for the moment that the power parameters α, β are known.

Sufficient information to estimate X, Y is given by the current allocation of resources in the region under study. If the current allocation pattern is described by x and y, Eqs. (7) and (8) may be rearranged as

$$X_{jk} = x_{jk}(\mu_{jk})^{1/(\alpha_j+1)} \qquad \forall j, k \tag{39}$$

$$Y_{jkl} = y_{jkl}\,(\lambda_l/C_l)^{1/(\beta_{jkl}+1)} \qquad \forall j, k, l \tag{40}$$

which are expressions for X and Y. We have a single equation for each unknown parameter, but as Section 3.4 predicted, we still need some external criterion to determine λ. If we assume that we can define the resources needed to satisfy the ideal levels X_{jk}, Y_{jkl} as some multiple θ_l of the resources used currently

$$\sum_j \sum_k X_{jk} Y_{jkl} = \theta_l \sum_j \sum_k x_{jk} y_{jkl} \qquad \forall l \tag{41}$$

then (39) and (40) can be substituted into (41) to give

$$f_l(\lambda) = 0 \qquad \forall l \tag{42}$$

where

$$f_l(\lambda) = -\theta_l \sum_j \sum_k x_{jk} y_{jkl} + \sum_j \sum_k x_{jk} y_{jkl} (\lambda_l)^{1/(\beta_{jkl}+1)} (\mu_{jk})^{1/(\alpha_j+1)} \tag{43}$$

and where Eq. (42) must be solved for λ. The equations in f are very similar to the equation $\hat{H}_\lambda = 0$ that arises during model solution, and, provided that $\theta_l > 1$, $\forall l$, and that all the terms except λ are known, they may be solved in the same way to give λ. Unfortunately, not all the terms are known. In particular, μ_{jk} is a weighted average involving the terms Y_{jkl}, which are as yet unknown. It is therefore necessary to iterate between solving Eq. (42) for λ, and Eqs. (39) and (40) for X, Y.

This approach suffers from the disadvantage that it only finds values of X, Y that are consistent with the current allocation pattern and the assumed values for α, β. A model with parameters estimated on so little data may have little predictive power. More useful is to estimate X, Y from other data and then to use the current allocation as a test of the model's validity. Other suitable data include cross-sectional and longitudinal data, and given N data points from such sources, we can use Eqs. (39) and (40) to find N estimates of X, Y. The problem remains of how to combine these estimates.

Estimates $X_{jk}(i)$, $Y_{jkl}(i)$ derived for subregions $i = 1, \ldots, N$, may be combined rather easily. If the population of subregion i is $P(i)$, then $X_{jk}(i)P(i)$ is the number of individuals in category j in mode of care k who need treatment in subregion i (per year), and $X_{jk}(i)Y_{jkl}(i)P(i)$ is the number of resources l needed to treat these individuals (per year). These quantities may be summed across the region, and the corresponding *regional* estimates of X, Y are

$$\bar{X}_{jk} = \sum_i X_{jk}(i)P(i) \Big/ \sum_i P(i) \qquad \forall j, k$$

$$\bar{Y}_{jkl} = \sum_i X_{jk}(i)Y_{jkl}(i)P(i) \Big/ \sum_i X_{jk}(i)P(i) \qquad \forall j, k, l$$

This approach (also depicted in Figure 10) is interesting because we do not need to assume that X, Y are constant across the region. The subregional variations are averaged by summing the ideal demands across the region.

Estimates $X_{jk}(i)$, $Y_{jkl}(i)$ derived at different times $i = 1, \ldots, N$ are more difficult to combine. Ideal supply levels Y_{jkl} are probably decreasing with time, and an exponential curve could be fitted to a long sequence of points. The ideal numbers of patients needing care per head of population, $Z_j = \Sigma_k X_{jk}$, $\forall j$, will change because of changes in the age structure and in the morbidity rates. We can correct for the former, but the latter are affected by changes in doctors'

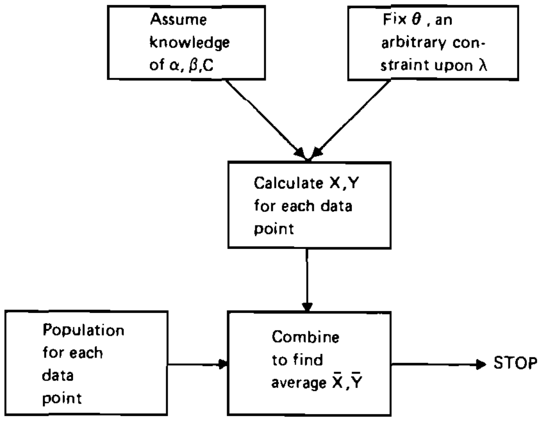


FIGURE 10 Estimation of ideal levels.

preferences between modes of health care. These are reflected in the values of X_{jk}, which could, if necessary, be regarded as experimental variables.

4.3 *Estimation of α, β*

We now consider how to estimate the power parameters α, β, assuming for the moment that the ideal levels X, Y are known. Sufficient information to estimate α, β is given by the current allocation of resources in the region under study. If the current allocation pattern is described by x and y, Eqs. (7) and (8) may be rearranged as

$$\alpha_j = \frac{\ln(\mu_{jk})}{\ln(X_{jk}/x_{jk})} - 1 \qquad \forall j, k \tag{44}$$

$$\beta_{jkl} = \frac{\ln(\lambda_l)}{\ln(Y_{jkl}/y_{jkl})} - 1 \qquad \forall j, k, l \tag{45}$$

which are expressions for α and β. As in Section 4.2, λ must be determined externally. We know, however, that α and β are always positive. This implies then that

$$\lambda_l > \tilde{\lambda}_l = \max_{j,k} \left[\left(\frac{X_{jk}}{x_{jk}} \right), \left(\frac{Y_{jkl}}{y_{jkl}} \right) \right] \qquad \forall l$$

and we can conveniently define λ_l as some (small) multiple $\phi_l > 1$ of the minimum value $\tilde{\lambda}_l$

$$\lambda_l = \phi_l \tilde{\lambda}_l \qquad \forall l \tag{46}$$

A second problem is that Eq. (44) gives K values for each α_j. Generally, these will be different values, but we can overcome this by aggregating the data across modes, and by using Eqs. (44) and (45) for one super mode.

By these means, we may estimate values for the parameters α, β. The model so calibrated will not exactly reproduce the current allocation of resources unless the latter is one of the admissible solutions of DRAM defined in Section 3. However, it will reproduce the actual supply levels y_{jkl}, and the actual numbers of patients in each category $(x_{j1} + x_{j2} + \ldots + x_{jK})$. Whether the estimated elasticities are useful for forward prediction will depend upon whether the current allocation pattern is representative of the HCS's usual behavior. The procedure described above only finds values for α, β that are consistent with this assumption and with the values assumed for X, Y.

A more sophisticated approach is to use more data by estimating *empirical elasticities*. These can then be used to derive the power parameters α, β. Appropriate empirical elasticities for DRAM are γ_{jkl}, the elasticity of the service level x_{jk} to changes in the resource level R_l, and η_{jkml}, the elasticity of the supply level y_{jkm} to changes in the resource level R_l. They can be predicted for given resource levels by DRAM. For example, γ_{jkl} is defined as

$$\gamma_{jkl} = \frac{\partial \ln x_{jk}}{\partial \ln R_l} = \frac{\partial \ln x_{jk}}{\partial \mu_{jk}} \frac{\partial \mu_{jk}}{\partial R_l} R_l$$

We use Eq. (7) to get an expression for $\partial \ln x_{jk}/\partial \mu_{jk}$. Thus,

$$\gamma_{jkl} = \frac{-R_l}{(\alpha_j + 1)\mu_{jk}} \frac{\partial \mu_{jk}}{\partial R_l} \tag{47}$$

Similarly,

$$\eta_{jkml} = \frac{-R_l}{(\beta_{jkl} + 1)\lambda_m} \frac{\partial \lambda_m}{\partial R_l} \tag{48}$$

where

$$\frac{\partial \mu_{jk}}{\partial R_l} = \sum_m \frac{\partial \mu_{jk}}{\partial \lambda_m} \frac{\partial \lambda_m}{\partial R_l}$$

and the derivatives $\partial R_l/\partial \lambda_m = \partial^2 H/\partial \lambda_l \partial \lambda_m$ are given by Eq. (13). Equations (47) and (48) can be written as

$$\alpha_j = \frac{A_{jkl}}{\gamma_{jkl}} - 1 \tag{49}$$

$$\beta_{jkm} = \frac{B_{ml}}{\eta_{jkml}} - 1 \tag{50}$$

where

$$A_{jkl} = \frac{-R_l}{\mu_{jk}} \sum_m \left(\frac{\partial \mu_{jk}}{\partial \lambda_m}\right) \bar{H}_{ml} \tag{51}$$

$$B_{ml} = \frac{-R_l}{\lambda_m} \bar{H}_{ml} \tag{52}$$

and where $\bar{H}_{ml}$ is element ml of the inverted Hessian matrix. However, solution for α, β is still hard. First, this is because A and B are functions of α and β, and iterative solution is necessary. Second, λ must still be chosen externally, and the empirical elasticities must be consistent with the choice of λ, otherwise the procedure may not converge (Gibbs 1978b). Third, there are more empirical elasticities γ, η than there are power parameters α, β. Therefore, unless some of the empirical elasticities are ignored, the parameters will be overspecified. Fourth, the empirical elasticities γ, η, are not directly measurable and are usually the result of some prior data analysis.

Some of these difficulties can be avoided by incorporating the prior data analysis within the solution of Eqs. (49)–(52). For example, estimates $\hat{\gamma}$, $\hat{\eta}$ are found by assuming that some N known data points $x_{jk}(i)$, $y_{jkl}(i)$, $R_l(i)$, $i = 1, \ldots, N$, satisfy the linear models

$$\ln x_{jk}(i) = a^x_{jk} + \sum_l \gamma_{jkl} \ln [R_l(i)] + \epsilon^x_{jk}(i) \tag{53}$$

$$\ln y_{jkm}(i) = a^y_{jkm} + \sum_l \eta_{jkml} \ln [R_l(i)] + \epsilon^y_{km}(i) \tag{54}$$

in which a^x, a^y are unknown constants, and ϵ^x, ϵ^y are random, uncorrelated error terms with zero means. If we eliminate γ, η by combining Eqs. (49), (50), (53), and (54) to give

$$\ln x_{jk}(i) = a^x_{jk} + \left(\frac{1}{\alpha_j + 1}\right) \sum_l A_{jkl} \ln [R_l(i)] + \epsilon^x_{jk}(i) \qquad \forall\, j, k, i \tag{55}$$

$$\ln y_{jkm}(i) = a^y_{jkm} + \left(\frac{1}{\beta_{jkm} + 1}\right) \sum_l B_{ml} \ln [R_l(i)] + \epsilon^y_{jkm}(i) \qquad \forall\, j, k, m, i \tag{56}$$

we can use the following iterative scheme in order to estimate α and β.

1. Fix λ arbitrarily for some resource level R, perhaps by using Eq. (46) on one of the data points.
2. Assume some initial estimates of α, β (e.g., unity).
3. Derive μ from Eqs. (9) and (10), $\hat{H}_{\lambda\lambda}$ from Eq. (13), and A, B from Eqs. (51) and (52).
4. Find the best least-squares estimators of $(\alpha_j + 1)^{-1}$, $(\beta_{jkm} + 1)^{-1}$ in Eqs. (55) and (56).
5. Hence, estimate α, β and repeat from step 3.

This procedure (also depicted in Figure 11) is likely to be lengthy because it incorporates regression estimation at each iteration. Nor can we ensure the positive estimates of α, β that are necessary for convergence. On the other hand, it has the advantage that more of the original data can be used directly. If a full data set

$$\{x_{jk}(i), y_{jkl}(i), R_l(i);\quad i = 1, \ldots, N,\ j = 1, \ldots, J$$
$$k = 1, \ldots, K,\ l = 1, \ldots, L\}$$

is available, KN equations are available to estimate each α_j, and perhaps not all of the $x_{jk}(i)$ need be known. Fewer equations (just N) are available to estimate each β_{jkl}, and it may be necessary to introduce some further simplifying

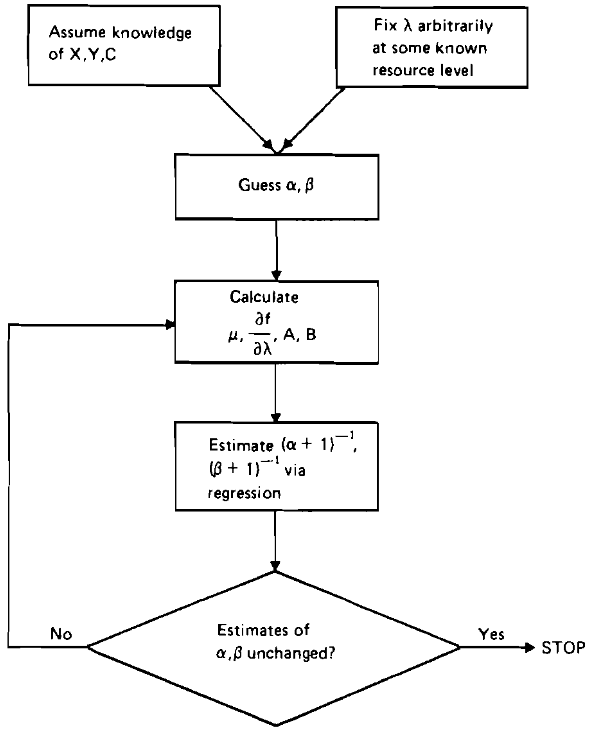


FIGURE 11 Estimation of power parameters.

assumptions such as $\beta_{j_1kl} = \beta_{j_2kl}$, $\forall j_1, j_2 \in \{1, \ldots, J\}$, in order to obtain reliable estimates. A second advantage of this procedure is that it is not necessary to modify any of the input data to make them consistent with the model. A third advantage is that the parameter estimated in each regression has an estimated standard error associated with it. These errors provide a measure of the reliability of α, β.

Perhaps the main assumption in the above analysis is that the underlying elasticities are constant across the set of data points. Because there is little information about how elasticities are likely to vary in time or space, we have not attempted to model this variation here. But Appendix D shows that in a certain sense, the procedure described above gives *unbiased* estimates. This is a reassuring result, and the estimates can be further tested to see if the model so calibrated can reproduce the current allocation of resources.

4.4 *Estimation of α, β and X, Y*

In the most general case, neither of the parameter pairs X, Y or α, β is known, and we require estimates of both. In this circumstance, the two procedures described above may be used together in the following scheme.

1. With some arbitrary initial estimates of α, β, use the methods of Section 4.2 to estimate X, Y.
2. With these estimates of X, Y, use the methods of Section 4.3 to estimate α, β.
3. Repeat from step 1.

The analysis in Section 3 showed that not even all small data sets can be consistent with DRAM, so that convergence of this scheme cannot be guaranteed. For this reason, although we have implemented on the computer the procedures for estimating both X, Y and α, β, we prefer not to link these programs together, but rather to use them alternately to obtain consistent pairs of estimates. (Note however that when neither parameter pair is given exogenously, the same data cannot be used to estimate both pairs of parameters.)

The parameter estimation procedures described above involve the choice of additional constraint variables such as ϕ and θ. Fortunately, however, this is not a problem. Although different values of ϕ, θ lead to different values for α, β, X, Y, each set of parameter values will reproduce with similar accuracy the data points used for estimation. Provided that predictive runs of the model do not involve resource levels very different from those used in estimation, the results will be relatively insensitive to ϕ, θ. Section 5 illustrates how these procedures were used to estimate model parameters in two examples.

4.5 *An Alternative Approach*

We now describe an alternative approach to parameter estimation that takes into account that DRAM's predictions are subject to uncertainty, and that incorporates this uncertainty mathematically. It is not fully implemented or tested, but the preliminary analysis given below is encouraging.

We consider how to use historical resource allocations $x(i)$, $y(i)$, $i = 1, \ldots, N$ in order to estimate the model parameter set $P = \{X, Y, \alpha, \beta\}$. As mentioned in Section 4.1, these are not the only data available. Nor does P include all the parameters: we have omitted the resource costs C because they seem to be more naturally estimated from external studies of financial or related statistics. Nevertheless, procedures to estimate these parameters from these data would be useful.

If reality conformed exactly to DRAM, we would expect the historical allocations $x(i)$, $y(i)$ to be exactly those $\hat{x}(i)$, $\hat{y}(i)$ prescribed by DRAM for the historical resource levels. These solutions are the result of (constrained) maximization over x and y of a function $U(x, y, P, C, R)$ that depends also upon the parameters P, the costs C, and the resource levels R. This function is known, and is presumably also maximized by choosing the correct parameters

$$\max_{P \text{ given past } \{x, y, R, C\}} U(x, y, P, R, C)$$

because with wrong parameters, it would be maximized by different values of x, y.

However, DRAM is only a model of reality. The historical allocations are related to the model predictions by equations like $x(i) = \hat{x}(i) + \xi_1(i)$ and $y(i) = \hat{y}(i) + \xi_2(i)$ where $\xi_1(i)$, $\xi_2(i)$ are stochastic processes with statistics S that need to be specified. Such a specification would be quite complicated. The probability distributions involved in S depend upon the reasons why the assumptions in DRAM are not perfect, the reasons that influence actual decisions, and the reasons that give rise to inaccurate data. But if such a specification were possible, the parameter set P could be estimated through

$$\max_{P} \underset{\substack{\text{with respect to} \\ \hat{x},\, \hat{y} \text{ given } x,\, y,\, S}}{\text{conditional expectation}} U(\hat{x}, \hat{y}, P, R, C) \tag{57}$$

where

$$U(\hat{x}, \hat{y}, P, R, C) = \max_{\substack{x, y \text{ given} \\ \text{past } \{R, C\}}} U(x, y, P, R, C) \tag{58}$$

Such a calculation would also be quite complicated, however, because the integral involved in the conditional expectation is unlikely to be analytic. In short, the ideal estimation procedure is extremely difficult both to formulate and solve. It does, however, suggest a more practical approach.

If the function $U(\hat{x}, \hat{y}, P, R, C)$ in Eq. (58) were twice differentiable in $\hat{x}$, $\hat{y}$, it could be expanded as a Taylor series about the point x, y, with terms in the prediction errors $(\hat{x} - x)$, $(\hat{y} - y)$. If, in addition, S were such that EXPECTATION $\xi_1(i)$ = EXPECTATION $\xi_2(i) = 0$, term-by-term expansion of this series would eliminate all first-order terms, causing the dominant terms of the series to be the squares and cross-products of the prediction errors. Whereas this is hardly a feasible way to solve (57), it suggests the idea of formulating the parameter estimation problem as the minimization of a function of the squared prediction errors

$$\min_{P} J(P) \tag{59}$$

where

$$J(P) = \frac{1}{2} \sum_{ijk} \rho^{x}_{ijk} [\hat{x}_{jk}(i) - x_{jk}(i)]^2 + \frac{1}{2} \sum_{ijkl} \rho^{y}_{ijkl} [\hat{y}_{jkl}(i) - y_{jkl}(i)]^2 \tag{60}$$

in which

$\hat{x}(i)$, $\hat{y}(i)$ are the optimal model allocations for assumed P and known past $R(i), C(i), i = 1, \ldots, N$

$x(i)$, $y(i)$ are the observed historical resource allocations for known past $R(i), i = 1, \ldots, N$

ρ^{x}_{ijk}, ρ^{y}_{ijkl} are weighting coefficients to be specified later.

DRAM's most useful feature is that the solutions $\hat{x}$, $\hat{y}$ are analytic functions of the parameters P. This means that we can calculate the gradient vector and Hessian matrix of $J(P)$, opening the way for powerful techniques for solving (59). The gradient vector is

$$\frac{\partial J(P)}{\partial P} = \sum_{ijk} \rho^x_{ijk}[\hat{x}_{jk}(i) - x_{jk}(i)] \frac{\partial \hat{x}_{jk}(i)}{\partial P} + \sum_{ijkl} \rho^y_{ijkl}[\hat{y}_{jkl}(i) - y_{jkl}(i)] \frac{\partial \hat{y}_{jkl}(i)}{\partial P} \quad (61)$$

and the Hessian matrix is

$$\frac{\partial^2 J(P)}{\partial P' \partial P} = \sum_{ijk} \rho^x_{ijk} \left\{ [\hat{x}_{jk}(i) - x_{jk}(i)] \frac{\partial^2 \hat{x}_{jk}(i)}{\partial P' \partial P} + \frac{\partial \hat{x}_{jk}(i)}{\partial P'} \frac{\partial \hat{x}_{jk}(i)}{\partial P} \right\}$$

$$+ \sum_{ijkl} \rho^y_{ijkl} \left\{ [\hat{y}_{jkl}(i) - y_{jkl}(i)] \frac{\partial^2 \hat{y}_{jkl}(i)}{\partial P' \partial P} + \frac{\partial \hat{y}_{jkl}(i)}{\partial P'} \frac{\partial \hat{y}_{jkl}(i)}{\partial P} \right\} \quad (62)$$

$$\simeq \sum_{ijk} \rho^x_{ijk} \frac{\partial \hat{x}_{jk}(i)}{\partial P'} \frac{\partial \hat{x}_{jk}(i)}{\partial P} + \sum_{ijkl} \rho^y_{ijkl} \frac{\partial \hat{y}_{jkl}(i)}{\partial P'} \frac{\partial \hat{y}_{jkl}(i)}{\partial P} \quad (63)$$

if the prediction errors are small. Expressions for the elements in the sensitivity derivative vectors $\partial \hat{x}_{jk}(i)/\partial P$ and $\partial \hat{y}_{jkl}(i)/\partial P$ are evaluated and listed in Appendix E.

The dimension of these vectors, and also of the Hessian matrix, is the same as the number of parameters $(2JKL + JK + J)$ in the parameter set P. Each element in the Hessian matrix is the sum of the $N(JK + JKL)$ terms enumerated in Eq. (63). Renumbering these terms as $m = 1, 2, \ldots, N(JK + JKL)$, we obtain the simpler form

$$\frac{\partial J(P)}{\partial P' \partial P} = \sum_m \rho_m v_m v'_m \quad (64)$$

where ρ are scalars

$$\rho_1 = \rho^x_{111}, \qquad \rho_2 = \rho^x_{112}, \ldots$$

$$\rho_{NJK+1} = \rho^y_{1111}, \qquad \rho_{NJK+2} = \rho^y_{1112}, \ldots$$

and v are vectors

$$v_1 = \frac{\partial \hat{x}_{11}(1)}{\partial P'}, \qquad v_2 = \frac{\partial \hat{x}_{12}(1)}{\partial P'}, \ldots$$

$$v_{NJK+1} = \frac{\partial \hat{y}_{111}(1)}{\partial P'}, \quad v_{NJK+2} = \frac{\partial \hat{y}_{112}(1)}{\partial P'}, \ldots$$

By arguments similar to those in Section 2.2, a matrix such as Eq. (64) is always positive semidefinite, which is useful for search procedures to solve (59).

However, the Hessian matrix will not be positive definite, and such searches will fail, unless the vectors v_m are linearly independent and span the parameter space. Just $2JKL + JK + J$ parameters X_{jk}, Y_{jkl}, α_j, β_{jkl}, $\forall j, k, l$, have to be estimated, and each data point x_{jk}, y_{jkl}, $\forall j, k, l$, provides $JK + JKL$ degrees of freedom that are subject to L resource constraints. Therefore, the number of data points N needed to identify P must satisfy $N(JK + JKL - L) \geqslant 2JKL + JK + J$. When $J = K = L = 1$, N must be four or more, but when $J = K = 3$ and $L = 2$, N can be as small as 2, although more data than this would be needed to achieve reasonable confidence in the estimated parameters.

An attempt to choose parameters P that will minimize $J(P)$ may also fail if the problem is badly conditioned, and specifically if the eigenvalues of

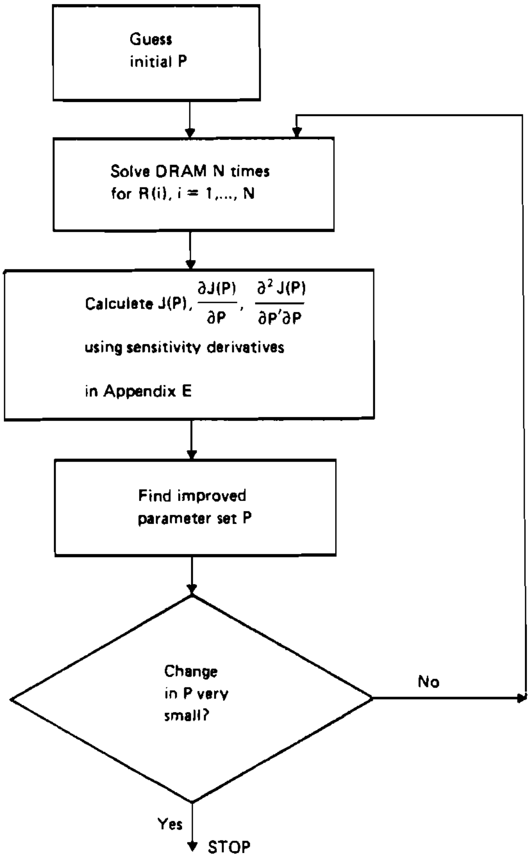


FIGURE 12 Proposed alternative parameter estimation procedure.

$\partial^2 J(P)/\partial P'\partial P$ are very dissimilar. We can control this by appropriate choice of the weights ρ in the definition of $J(P)$ in Eq. (62). Setting $\rho_m = \|v_m\|^{-2}$ is equivalent to normalizing all the vectors in Eq. (64) to unit length. If the vectors are additionally orthogonal, all the eigenvalues would be equal. When they are not orthogonal, the eigenvalues are approximately equal.

Figure 12 shows a way of using these results to estimate the parameter set P by solving (59) according to an iterative procedure. It uses initially some guesses about P to derive the function $J(P)$ in Eq. (60), and then the gradient vector (61) and perhaps the Hessian matrix (62) to find a new parameter set that is closer to the solution of (59).

What computation is involved in this procedure? At each step in the iteration, DRAM must be solved N times to give the model predictions $x(i)$, $y(i)$ corresponding to each of the observed data points $x(i)$, $y(i)$. Probably this procedure is most useful when large amounts of data are available (N at least greater than 20). This means more than 20 DRAM solutions per step, and probably at least 200 DRAM solutions for convergence. But with the typical model solution times reported in Appendix B, this is not too many, especially when each solution also gives the gradient vector and Hessian matrix of $J(P)$.

Potentially, the storage requirements could be excessive. Fortunately, however, all three terms $J(P)$, $\partial J(P)/\partial P$, $\partial^2 J(P)/\partial P' \partial P$ are formed by summation, and the individual terms can be calculated and added sequentially. Appendix E shows that many of the sensitivity derivatives are identically zero, and the remaining derivatives can be computed in logical and space-saving order. The Hessian is symmetrical, permitting further saving. For Example 2 in Section 5, where $J = 7$, $K = 2$, and $L = 2$, the number of locations needed to store these three functions is $1 + (2JKL + JK + J) + \frac{1}{2}(2JKL + JK + J)(2JKL + JK + J + 1) = 1 + 77 + 3003 = 3081$, which is quite reasonable. It remains only to specify how a new parameter set is determined. This problem is similar to that of finding improved estimates of λ in Section 2.4, and similar or more sophisticated gradient methods can easily be devised.

4.6 *Estimation of C*

We now discuss how to estimate the unit resource costs C needed in the model. These parameters are defined rather carefully. Specifically, C_l is the marginal cost of using one more resource of type l, when all needs for health care are met. Strictly speaking, these costs are not money costs but opportunity costs. They reflect the benefit in some alternative that is foregone through buying the extra resource. How then can they be estimated? Often, we have financial data that we can use directly, but when these are unavailable or inappropriate, how can equivalent model parameters be inferred?

Two assumptions will enable us to estimate the costs C from financial data, when these are available. The first assumption is that in long-term planning, opportunity costs are approximately measured by money costs. Given sufficient time, every option is an alternative, and all resources are substitutable. The second assumption is that marginal costs are approximately measured by average costs. The cost function of an individual hospital or medical school is certainly nonlinear, with marginal costs being generally less than average costs. But when many such hospitals or medical schools are operating in a single region, the aggregate cost function may be approximately linear, as shown in Figure 13. In these circumstances, the average costs recorded in historical accounts approximate the marginal costs at some hypothetical resource level.

However, not all countries compare alternative plans in terms of financial feasibility. In the Soviet Union, for example, planning seeks mainly to reconcile the real outputs between producers while satisfying aims such as full employment and constant growth. For application of the model in these countries, it is not necessary to estimate resource costs, but only some parameters that have an equivalent function in the model. The purpose of the C parameters is to

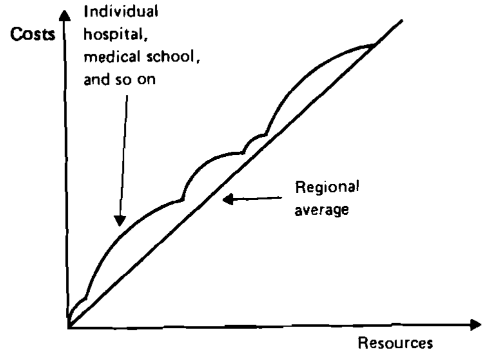


FIGURE 13 A linear regional cost function.

reflect the relative value of different resources; or, conversely, their relative scarcity or the relative difficulty of providing different resources. In a society with uniform and constant growth, resources increase in proportion to their current levels, and these ratios may be adequate first estimates of the C parameters. When different growths are expected in different parts of the HCS, the ratios may be adjusted accordingly, or more detailed analysis may reveal the shadow prices of each constrained resource.

The principal unsolved problem is that of resource definition. The costs of a hospital bed could be the capital cost of creating it, or the cost of maintaining the patient in it with food, heat, and laundry. The cost of a doctor could include his training, his accommodation, or just his salary. The choices made at this stage actually *define* the resources for the purposes of the model, and they depend mainly upon which alternatives are interesting to the users of the model. Finally, of course, we desire to estimate C at some future time instead of at the present. A full treatment of this issue would need and could use more sophisticated predictive models.

5 EXAMPLES

This section contains no mathematics; instead, it illustrates how DRAM can be used. Section 5.1 has some general comments about the use of mathematical models, and Sections 5.2 and 5.3 contain two examples. The first is used mainly to compare different methods of parameter estimation. The second uses the full structure available in DRAM (categories, modes, and resources) in order to investigate questions of resource balance.

5.1 *Application of DRAM*

A mathematical model represents some common mechanism or process. The process that DRAM represents is the distribution of scarce resources within a large and complex system. The value of having such a model is that changes proposed for the system can be tried out first on the model to see what effects they are likely to have. This helps in debates about which changes are best.

Three points need to be noted. First, DRAM is not a model of a complete HCS. Rather, it is a model of resource allocation in such systems. Second, DRAM is not a model of health resource allocation in, for example, Austria. Rather, it is a model for all regions (nations or districts) where its hypotheses are justified. Third, DRAM is not a model with certain specific data needs. Rather, it is a tool that can be calibrated for different problems, large and small.

A large problem might concern the use of all health resources throughout a country. To apply DRAM here (we have not attempted it), we would need a detailed study of the appropriate patient categories and resource groupings. Because DRAM uses generalized variables (e.g., resources) which are not restricted in number or type, as many categories as desired, of whichever type, could be used. A lot of data about past allocation patterns would need to be collected and related to other sources in order to estimate parameters, and methods such as those proposed in Section 4.5 would be useful. Such a model, linked with other models for population, morbidity, and education, would be a tool similar in scope to a large-scale economic model.

A small problem might concern the age distribution of hospital patients. In such an exploratory application, not all the dimensions available in DRAM would be needed. Just three patient categories (young, middle-aged, and old), one resource (beds) and one mode (hospital) might be enough. But if subsequent work suggests that lack of convalescent care is affecting discharges from hospital, then no new structure would be needed to extend the analysis to include this extra mode of care. It might be interesting to use the model with alternative age groupings to see if the results are sensitive to this. Because DRAM is easy to solve, many runs are possible at small expense.

What sort of problems are amenable to investigation with DRAM? The most obvious ones are questions about the consequences of changing levels of resources. When all resources are increasing, DRAM probably has little to say about who gets what, beyond what could be deduced directly from the empirical elasticities of demand to supply. But when some resources are increasing (e.g., numbers of doctors), and others are decreasing (e.g., hospital beds), either by design or through natural trends, such simple deductions become difficult. DRAM recognizes that there will be substitution between resources, and can show where the balance will lie.

Slightly different questions arise when resource levels are constant but the behavior of the HCS is changing. Morbidity levels (the X parameters in DRAM) change with population age structure. Ideal standards of care (the Y parameters in DRAM) change as alternative forms of care become popular. These sorts of assumptions lead to model runs that predict what will happen in the future to a single sector (e.g., care of children), if no change is made in the present HCS.

More unorthodox applications are possible. DRAM is deliberately designed with parameters that can be interpreted outside the model. When ideal standards of care (Y) have been proposed by professional consensus, DRAM is useful for seeing how nearly they can be achieved when resources are scarce. But this approach can also be reversed. The parameter estimation procedures reveal what ideal standards are implied by current behavior, and how these compare with professionally set standards. Such procedures can also be used to estimate the levels of potential demands for care, and thereby make a comparison of underlying morbidity. Effectively, the model is inverted in order to predict inputs from outputs. The examples that follow, however, are rather more straightforward.

5.2 *Example 1: Hospital Beds*

If more hospital beds are provided to increase the numbers of short-stay patients, might the result just be the same number of long-stay patients staying still longer? Because hospital beds are an expensive form of care, this is an important question. To illustrate how DRAM can be used to study it, consider the distribution of acute hospital bed-days between patients suffering from six diseases: varicose veins, hemorrhoids, ischemic heart disease (excluding acute myocardial infarction), pneumonia, bronchitis, and appendicitis. Table 1 gives the numbers of patients admitted to hospitals in England in 1968 and 1973 with these diseases, and their average lengths of stay (Department of Health and Social Security 1972, 1977a). Together, these patients use only about 8 percent of all hospital beds (excluding maternity beds), but an extension of this example to include the remainder, either as a group or individually, would not be difficult. We notice that during these 5 years, the number of bed-days used for these diseases has fallen by about 28 percent. Furthermore, admissions and lengths of stay in each disease category have nearly all fallen. Is it possible to calibrate a model of these changes?

Gibbs (1977, 1978a, b) did this using the empirical elasticities estimated by Feldstein (1967) from 1960 data, and exogenous 1968 estimates of the ideal levels X, Y. The corresponding model parameters, summarized in Table 2, were used to reproduce the 1968 allocations in one region of England (the South Western Regional Health Authority – SWRHA), and to investigate the effects of changing the number of beds available there by 20 percent. The analysis was repeated with X, Y chosen to reproduce regional admission and supply levels.

We have repeated this exercise, applying the parameter estimation methods described in Section 4 to the actual admissions and lengths of stay in the 14 health regions of England in 1968 and 1973 (Department of Health and Social Security 1972, 1977a). Table 3 gives the parameters estimated by using the 1968 figures to estimate α, β and the 1973 figures to estimate X, Y recursively

TABLE 1 Allocation of hospital bed-days in England.

	1968		1973	
	Admissions per 10,000 people	Average stay (days)	Admissions per 10,000 people	Average stay (days)
Varicose veins	9.8	12.0	7.6	10.1
Hemorrhoids	5.6	10.1	4.7	7.8
Ischemic heart	6.5	39.8	8.5	24.9
Pneumonia	14.2	25.4	14.0	18.0
Bronchitis	14.1	25.6	10.8	23.1
Appendicitis	20.4	9.1	17.5	7.9
Total bed-days per 10,000 people	1,340.1		964.8	

SOURCE Department of Health and Social Security (1972, 1977a).

TABLE 2 First set of model parameters for Example 1.

	Empirical elasticities[a]		Model parameters			
	γ	η	α^b	β	X^b	Y
Varicose veins	0.78	0.62	1.64	3.03	12.8	15.4
Hemorrhoids	0.70	0.44	2.11	4.68	7.7	13.1
Ischemic heart	1.14	1.08	0.54	1.31	10.4	52.1
Pneumonia	0.71	0.23	2.28	9.87	21.0	19.7
Bronchitis	1.13	−0.23	1.14	49.00	21.3	34.2
Appendicitis	−0.16	0.31	44.40	7.06	24.8	10.1

[a] Feldstein (1967, p. 219).
[b] α was estimated from γ, η with arbitrary constant $c = 25$, and X was chosen exogenously (Gibbs 1978b).

as described in Section 4.4. For this example, we have assumed that the parameters are constant over time, but we could have incorporated exogenous information to correct for this. We could also have corrected for the effects of changing age structure, but they were small. At some points in the iteration towards the results of Table 3, negative elasticities were estimated, but their associated standard errors were so large that they could reasonably be changed to small positive numbers. If professional opinions about ideal admission rates or lengths of stay had been available to us, we could have incorporated them also within this scheme.

Gibbs (1978b) used data from the SWRHA for model testing, and we have done the same. In 1973, only 633 bed-days per 10,000 people were used for the six diseases and Table 4 shows how they were distributed. Making the

TABLE 3 Second set of model parameters for Example 1.

	Empirical elasticities[a]		Model parameters[b]			
	γ	η	α^c	β	X	Y
Varicose veins	0.54	0.43	1.68 (0.7)	3.27 (0.5)	31.6	30.9
Hemorrhoids	0.34	0.31	3.63 (0.5)	5.00 (−0.9)	11.2	17.0
Ischemic heart	0.66	0.93	0.50 (0.7)	1.00 (−12)	71.0	247.5
Pneumonia	0.66	0.18	1.57 (0.8)	9.44 (0.5)	75.2	28.1
Bronchitis	0.90	0.04	1.04 (0.8)	50.00 (−4)	102.7	24.9
Appendicitis	0.04	0.14	40.00 (0.3)	12.75 (0.6)	19.5	11.1

[a]Derived from α, β, X, Y and R = 1340.1 bed-days per 10,000 people.
[b]Estimated from 1968 and 1973 allocations across 14 English regions (Department of Health and Social Security 1972, 1977a) with arbitrary constants $\phi = 5$, $\theta = 20$.
[c]Confidence coefficients (in parentheses) are defined as 1 − (estimated standard error ÷ estimated value).

TABLE 4 Allocation of hospital bed-days[a] in 1973 in the South Western Region of England (Example 1).

	Actual[b]		Predicted by model using Table 2 parameters		Predicted by model using Table 3 parameters	
	Admissions per 10,000 people	Average stay (days)	Admissions per 10,000 people	Average stay (days)	Admissions per 10,000 people	Average stay (days)
Varicose veins	6.1	14.4	5.5	8.1	6.1	8.7
Hemorrhoids	4.2	7.7	3.7	8.3	4.1	6.9
Ischemic heart	5.3	17.4	3.0	16.9	6.5	16.5
Pneumonia	11.4	14.4	9.9	15.5	10.7	16.7
Bronchitis	9.9	16.8	6.5	32.5	7.5	22.4
Appendicitis	15.4	7.8	23.5	7.3	17.2	7.5

[a]663 bed-days available per 10,000 people in 1973.
[b]From Department of Health and Social Security (1977a).

assumption that the model parameters that we have estimated from English data are appropriate to SWRHA, we can use the model to make predictions of this distribution, also shown on Table 4. The parameters from Table 3 give slightly better predictions than those from Table 2; the average error is about

14 percent. Note also that predictions from two sets of model parameters indicate the sensitivities of the model outputs to changes in model parameters. (Appendix E shows that expressions for these sensitivities can also be derived explicitly.) If these parameters are judged acceptable, the model can be used with different bed supply levels to predict the effects of an increase or a decrease in the number of beds. It is important to note that such predictions have little value unless the model is adequately calibrated. It is for this reason that parameter sets estimated from different sources are valuable.

The two sets of model parameters in Tables 2 and 3 vary because of different data and because of different values used for the arbitrary constants. (It is not easy to choose equivalent values when both procedures are solved iteratively.) Nevertheless, they show very similar variations across diseases. Appendicitis is clearly represented as a disease where most patients go to the hospital (high α), and bronchitis appears as a disease afflicting many patients (high X) for whom hospital care is not essential (low α). The empirical elasticities in Table 3 are values derived via the model, Eqs. (47) and (48), using the 1968 English resource level. This calculation incorporates DRAM's behavioral assumption (see Section 1). Because Feldstein's estimates given in Table 2 do not incorporate this assumption, the reasonable agreement between them suggests that DRAM's assumptions are valid, and supports the previous results.

5.3 *Example 2: The Balance of Inpatient and Outpatient Care*

If hospital beds are decreased and medical staff are increased, will more or fewer patients receive treatment and how will the balance of inpatient and outpatient care be affected? This is a question facing health managers in England and elsewhere, and DRAM can be used to help answer it.

Table 5 shows how beds and doctors were used in the SWRHA in England for 1977 in the seven largest acute hospital specialties: general surgery, general medicine, obstetrics and gynecology, trauma and orthopedic (T & O) surgery, ear, nose, and throat (ENT), pediatrics, and ophthalmology (Department of Health and Social Security 1977b). In this example, the patient categories are the seven specialties, the two modes of care are inpatient and outpatient, and the two resources are beds and doctors. Therefore, this example uses all the structure available in DRAM, although it has the simplifying feature that one of the resources (beds) is used in only one mode of care (inpatient).

Because the problem is more complicated than the previous one, formulating a suitable DRAM model is more difficult. For example, hospital specialties are not as precisely defined as disease categories, and the division of doctor's time between inpatients and outpatients is not directly measurable. The first is not so important if the definitions are reasonably consistent across the region. If the definitions are consistent but not universal, comparisons beyond SWRHA may be suspect. The second difficulty can be overcome by subtracting from each consultant's working year (measured in half days), the number of

TABLE 5 Beds and doctors in the South Western Regional Health Authority in 1977 (Example 2).

	Relevant catchment population (thousands)	Admissions per 1,000 people		Average hospital stay (days)[a]	Half-day consultant sessions per admission[b]	
		Inpatient	Outpatient		Inpatient[c]	Outpatient
General surgery[d]	3,035.4	20.9	19.0	7.87	0.170	0.153
General medicine[e]	3,035.4	14.8	10.5	10.18	0.183	0.345
Obstetrics and gynecology	1,563.8[f]	39.5	37.1	5.78	0.072	0.139
T & O surgery	3,035.4	9.1	22.4	13.60	0.252	0.121
ENT	3,035.4	4.4	11.1	4.39	0.346	0.128
Pediatrics	641.8[g]	29.7	17.7	6.28	0.266	0.362
Ophthalmology	3,035.4	2.8	10.3	6.59	0.427	0.214

SOURCE Department of Health and Social Security (1977b).

[a] 892 bed-days available per 1,000 people in 1977.

[b] Assuming each full-time consultant works the equivalent of 450 half-day sessions per year; 46 half-day consultant sessions available per 1,000 people in 1977.

[c] Derived by subtracting actual outpatient sessions from total number of sessions.

[d] Includes urology.

[e] Includes cardiology.

[f] Excludes males.

[g] Excludes people more than 15 years old.

TABLE 6 Estimated model parameters for Example 2.

	α_j	X_{j1} (Inpatient)	X_{j2} (Outpatient)	β_{j11} (IP, beds)	β_{j12} (IP, doctors)	β_{j22} (OP, doctors)	Y_{j11} (IP, beds)	Y_{j12} (IP, doctors)	Y_{j22} (OP, doctors)
General surgery[a]	10.0	26.3	22.2	10.8	6.1	1.0	10.5	0.34	0.46
	(−0.4)			(0.7)	(0.7)	(−3.9)			
General medicine[b]	0.01	217.7	83.3	10.7	2.7	11.2	13.3	0.42	0.41
	(0.5)			(0.3)	(0.7)	(0.8)			
Obstetrics and gynecology	16.5	44.8	38.7	10.3	1.5	0.001	7.7	0.22	1.32
	(0.2)			(0.6)	(0.8)	(0.6)			
T & O surgery	10.0	10.8	26.5	1.0	12.7	10.0	58.5	0.37	0.15
	(−1.4)			(−1.4)	(0.2)	(−10.7)			
ENT	10.0	5.0	12.9	0.001	14.3	20.0	79.1	0.43	0.15
	(−0.7)			(0.2)	(0.0)	(0.7)			
Pediatrics	5.6	43.7	19.4	8.9	5.8	1.0	9.1	0.41	1.28
	(0.7)			(0.4)	(0.7)	(−4.8)			
Ophthalmology	20.0	3.1	11.9	10.0	8.3	10.0	9.4	0.60	0.24
	(0.0)			(−2.8)	(0.3)	(−2.3)			

NOTE: Confidence coefficients as defined in Table 3 appear in parentheses.

[a] Includes urology.

[b] Includes cardiology.

TABLE 7 Validation results for Example 2.

	Relevant catchment population (thousands)	Admissions per 1,000 people		Average hospital stay (days)[a]	Half-day consultant sessions per admission[b]	
		Inpatient	Outpatient		Inpatient[c]	Outpatient
Actual resource allocation in SWRHA in 1975[h]						
General surgery[d]	3,003.7	19.6	16.7	8.54	0.253	0.166
General medicine[e]	3,003.7	14.3	8.5	10.91	0.252	0.372
Obstetrics and gynecology	1,555.8[f]	35.8	31.1	6.08	0.115	0.147
T & O surgery	3,003.7	8.3	18.9	14.25	0.246	0.139
ENT	3,003.7	4.2	9.3	4.46	0.405	0.152
Pediatrics	654.7[g]	31.0	14.2	7.20	0.279	0.398
Ophthalmology	3,003.7	2.7	10.3	7.44	0.533	0.194
Predicted resource allocation in SWRHA in 1975						
General surgery[d]		20.6	18.9	8.22	0.244	0.138
General medicine[e]		15.1	8.4	10.41	0.220	0.335
Obstetrics and gynecology		38.5	36.1	5.96	0.082	0.118
T & O surgery		9.0	21.5	14.05	0.314	0.120
ENT		4.5	10.4	4.58	0.365	0.136
Pediatrics		29.4	14.9	6.83	0.285	0.381
Ophthalmology		2.7	10.7	7.24	0.463	0.194

[a] 922 bed-days available per 1,000 people in 1975.
[b] Assuming each full-time consultant works the equivalent of 450 half-day sessions per year; 48 half-day consultant sessions available per 1,000 people in 1975.
[c] Derived by subtracting actual outpatient sessions from total number of sessions.
[d] Includes urology.
[e] Includes cardiology.
[f] Excludes males.
[g] Excludes people more than 15 years old.
[h] From Department of Health and Social Security (1977b).

outpatient sessions worked during that year in that specialty. The ratio of the cost of a doctor to the cost of a bed is assumed to be 1.57: 1 (Hughes 1978a). In deriving this figure, the cost of each bed includes all associated costs *except* the cost of the doctor.

Table 6 shows the model parameters that were estimated by the methods of Section 4 from historical allocation data from 1976 and 1977, and disaggregated for the five hospital areas of the SWRHA. With only ten data points we would not expect to estimate a complete parameter set with great confidence, and some of the figures in Table 6 are very uncertain. Nevertheless, the variations between parameters are as expected. In obstetrics and gynecology most of the demand is met (high α_j) but the need for outpatient treatment is very elastic (low β_{j22}). In general medicine, the reverse is true. Many patients do not receive hospital care, but the supply of resources to those who do is rather inelastic.

Table 7 compares the predictions made by the model using these parameters with the actual allocations in 1975. The agreement between model and reality is better than that found in Example 1, but this is partly because of relatively small changes in the SWRHA during the 3 years. Further calibration tests would be desirable.

Meanwhile, however, we consider how to use this model to answer the question at the beginning of this section. We want to increase the numbers of doctors, but this can be afforded only by decreasing the number of beds. We imagine that, from the 1975 resource levels, doctors are increased by 10 percent and beds decreased by 10 percent. (With only tentative parameter estimates, predictions for larger changes may be suspect.) What will happen? The response of the HCS could be to

- Treat different numbers of patients
- Use more or fewer resources per patient
- Change the specialty mix of patients treated
- Change the mix of resources used to treat different patients
- Change the mode of treatment between inpatient and outpatient care for different patients

The simple proportional changes do not indicate which effect will dominate; the model can.

Table 8 shows the predicted results of decreasing beds and increasing doctors, each by 10 percent. As might be expected, these changes result in fewer inpatients and more outpatients. Because of the several population divisors, the total percentage shifts are difficult to quantify, but inpatients decline by about 8 percent, and outpatients increase by about 6 percent. The remaining changes take place in the average lengths of stay and in the distribution of doctor's time among patients.

It is interesting to examine whether, when inpatients and outpatients are

TABLE 8 Predicted results for a decrease in beds and an increase in doctors (Example 2).

	Admissions per 1,000 people		Average hospital stay (days)[a]	Half-day consultant sessions per admission[b]	
	Inpatient	Outpatient		Inpatient[c]	Outpatient
General surgery[d]	20.1	19.2	8.02	0.255	0.161
General medicine[e]	11.7	11.2	10.15	0.240	0.344
Obstetrics and gynecology[f]	38.0	36.3	5.81	0.093	0.161
T & O surgery	8.8	22.1	12.15	0.321	0.123
ENT	4.4	10.7	3.43	0.373	0.138
Pediatrics[g]	28.3	15.4	6.63	0.298	0.445
Ophthalmology	2.7	10.9	7.05	0.479	0.200

[a] 830 bed-days available per 1,000 people (10 percent less than in 1975).
[b] Assuming each full-time consultant works the equivalent of 450 half-day sessions per year; 52 half-day consultant sessions available per 1,000 people (10 percent more than in 1975).
[c] Derived by subtracting actual outpatient sessions from total number of sessions.
[d] Includes urology.
[e] Includes cardiology.
[f] Excludes males.
[g] Excludes people more than 15 years old.

added together, more or fewer patients are treated in each specialty. The model suggests increases in T & O surgery, ENT, and pediatrics, and decreases in the other specialties. The specialty with the largest change from inpatient to outpatient care is general medicine. Naturally, all the lengths of stay decrease, most notably in T & O surgery (by 2 days) and ENT (by 1 day). Naturally, all the levels of doctor care rise, but some of them hardly at all (e.g., T & O surgery and ENT). The largest increases occur in obstetrics and gynecology, with the implication that doctors are under most pressure in these specialties.

Of course, a decision about changing resource levels may be more complicated than represented above. In England, for example, approval for new consultant posts is granted in specific specialties. But a model run in which total consultant posts are increased is still useful in suggesting the specialties for which approval should be sought. The response of the system is also likely to be more complicated than represented above. For example, utilization measures such as bed occupancy may change, thereby upsetting DRAM's predictions. If this happens, a model of the more critical resources may be more appropriate, and DRAM is sufficiently flexible to allow this. Whenever data from past years are available which show how resources were distributed between categories and modes, such data can be used to test DRAM's hypothesis and, if possible, to calibrate a relevant model.

6 SUMMARY

Health care systems are unlike the more common engineering systems that are investigated by mathematical modelers. They are social systems, inaccessible for experiment, where many different agents act according to personal preferences, and without any operational definition of the principal output – health. The chances of using mathematical analysis to study resource allocation would seem to be slight. How then have we done so much algebra?

In fact, nearly all the algebra derives from just two equations – Eq. (5), which says that all resources are used, and Eq. (1), which says that the system tries to give the most care to the most people. Section 2 showed how these two equations are sufficient to derive Eqs. (7) and (8), which say which individuals get what sort of care. These equations constitute DRAM, and the rest of the report looks at the results that they predict.

The predictions will be good ones only if the two underlying equations are realistic. Because justification by common sense can be wrong, we have investigated in Section 3 the sorts of resource allocation patterns that DRAM can imply. This analysis found that the model cannot reproduce increasing levels of service and decreasing levels of supply simultaneously, but that it will always make use of all the available modes of care. Such results make DRAM applicable in many different sectors of health care, and perhaps elsewhere.

For DRAM to be useful, it must be possible to put numbers into the equations on the basis of observed data. Section 4 presented methods that use routine statistics, but that take into account that all sources of data may reflect inherent parameter variations. It is also possible to put numbers into the model on intuitive or professional advice, and some of our procedures indicate which of the parameters might be improved by intelligent guesswork.

Practical application of the model requires cheap and speedy solutions. The computing times reported in Appendix B indicate a very efficient solution algorithm. Even a program with full error handling and diagnostics is still quite small and easy to install.

For what purposes can we use DRAM? Section 5 discussed large and small applications, and two problems amenable to DRAM were investigated in two examples. The first was concerned with allocation of beds among patients with different diseases. The second dealt with the question: Will more or fewer individuals be treated in South Western England if hospital beds are decreased by 10 percent and hospital consultants are increased by 10 percent? The answer (more in some specialties, fewer in others) could be the beginning of a more detailed analysis.

Questions like these are not easy to answer from tables of statistics alone, and DRAM can be seen as a way of organizing information to help in problems of resource allocation. Section 4 therefore examined ways to make DRAM easier to set up when a lot of data are available. These methods are attractive because they derive from an ideal approach to estimating parameters, yet seem feasible and even efficient. Testing them within case study applications is a task for the future.

ACKNOWLEDGMENTS

We are indebted to many colleagues at IIASA and elsewhere for their help and advice. We would like to offer special thanks to Rebecca Crow, who prepared the typescript, tables, and diagrams, and to the staff of the South Western Regional Health Authority, who provided the data used in Section 5.

REFERENCES

Burton, R. M., D. C. Dellinger, W. M. Damon, and E. A. Pfeiffer (1978) A Role for Operational Research in Health Care Planning and Management Teams. Journal of the Operational Research Society 29 (7): 633–641.

Cardus, D. and R. M. Thrall (1977) The Concept of Positive Health and the Planning of Health Care Systems. In D. D. Venedictov, ed., Health System Modeling and the Information System for the Coordination of Research in Oncology. Proceedings of the IIASA Biomedical Conference. CP-77-4. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Department of Health and Social Security, U.K. (1972) Report on Hospital In-Patient Enquiry for the Year 1968. London: Her Majesty's Stationery Office.

Department of Health and Social Security, U.K. (1977a) Report on Hospital In-Patient Enquiry for the Year 1973. London: Her Majesty's Stationery Office.

Department of Health and Social Security, U.K. (1977b) Hospital Medical Staff Tables, SH3 Hospital Return Summaries, 1975, 1976 and 1977. Internal reports.

Feldstein, M. S. (1967) Economic Analysis for Health Service Efficiency. Amsterdam: North-Holland.

Fujimasa, I., S. Kaihara, and K. Atsumi (1978) A Morbidity Submodel of Infectious Diseases. RM-78-10. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Gibbs, R. J. (1977) Health Care Resource Allocation Models – A Critical Review. RM-77-53. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Gibbs, R. J. (1978a) A Disaggregated Health Care Resource Allocation Model. RM-78-1. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Gibbs, R. J. (1978b) The IIASA Health Care Resource Allocation Sub-Model: Mark 1. RR-78-8. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Hughes, D. J. (1978a) The IIASA Health Care Resource Allocation Sub-Model: Mark 2 – The Allocation of Many Different Resources. RM-78-50. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Hughes, D. J. (1978b) The IIASA Health Care Resource Allocation Sub-Model: Formulation of DRAM Mark 3. WP-78-46. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Hughes, D. J., E. A. Nurminski, and G. Royston (1979) Nondifferentiable Optimization

Promotes Health Care. WP-79-90. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Kaihara, S., I. Fujimasa, K. Atsumi, and A. A. Klementiev (1977) An Approach to Building a Universal Health Care Model: Morbidity Model of Degenerative Diseases. RM-77-6. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Lemarechal, C. and R. Mifflin, eds. (1978) Nonsmooth Optimization. Proceedings of a IIASA Workshop. Oxford: Pergamon.

McDonald, A. G., G. C. Cuddeford, and E. M. L. Beale (1974) Mathematical Models of the Balance of Care. British Medical Bulletin 30(3): 262–270.

Nackel, J. G., J. Goldman, and W. L. Fairman (1978) A Group Decision Process for Resource Allocation in the Health Setting. Management Services 24 (12): 1259–1267.

Newhouse, J. P. and C. E. Phelps (1974) Price and Income Elasticities for Medical Care Services. R-1197-NC/OEO. Santa Monica, California: The Rand Corporation.

Rousseau, J. (1977) The Need for an Equilibrium Model for Health Care System Planning. In E. N. Shigan and R. Gibbs, eds., Modeling Health Care Systems – Proceedings of a IIASA Workshop. CP-77-8. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Shigan, E. N., D. J. Hughes, and P. I. Kitsul (1979) Health Care Systems Modeling at IIASA: A Status Report. SR-79-4. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Van der Gaag, J., F. Rutten, and B. van Praag (1975) Determinants of Hospital Utilization in the Netherlands. Health Services Research 10 (3): 264–278.

Venedictov, D. D., A. A. Blyusin, T. G. Berezneva, K. M. Kelmanson, A. S. Kiselev, A. A. Klementiev, Yu. Komarov, V. N. Novoseltsev, A. M. Petrovsky, M. A. Schnepps-Schneppe, E. N. Shigan, P. P. Volkov, and A. I. Yashin (1977) Health Care: A Systems Approach. In D. D. Venedictov, ed., Health System Modeling and the Information System for the Coordination of Research in Oncology. CP-77-4. Laxenburg, Austria: International Institute for Applied Systems Analysis.

APPENDIXES

A AN ALTERNATIVE FORMULATION OF DRAM

The formulation of the DRAM model depends upon the definition in Eq. (1) of the function $U(x, y)$. An alternative definition

$$U(x, y, z) = \sum_j g_j(z_j) + \sum_j \sum_k \sum_l x_{jk} h_{jkl}(y_{jkl}) \tag{A1}$$

was investigated by Hughes (1978b), in which

$$g_j(z) = \frac{c_j Z_j}{\alpha_j} \left[1 - \left(\frac{z}{Z_j} \right)^{-\alpha_j} \right] \tag{A2}$$

$$h_{jkl}(y) = \frac{C_l Y_{jkl}}{\beta_{jkl}} \left[1 - \left(\frac{y}{Y_{jkl}} \right)^{-\beta_{jkl}} \right] \tag{A3}$$

and where the HCS is assumed to want to increase the total number of individuals in category j who receive care (per head of population per year)

$$z_j = \sum_k x_{jk} \qquad \forall j \tag{A4}$$

irrespective of the numbers x_{jk} in each mode of care. The parameters Z_j represent the total number of individuals in category j who need care. The parameters c_j are the marginal costs of treating one individual in category j, when all demands are met

$$c_j = \sum_k \sum_l C_l Y_{jkl}$$

The other parameters are as defined for the original DRAM.

An important property of $U(x, y)$ in the original DRAM is that $x_{jk} \to 0$ for any j, k, causes $U \to -\infty$. Because the solution to DRAM maximizes U, this condition automatically excludes solutions in which any $\hat{x}_{jk}$ is zero. However, this condition is not true of $U(x, y, z)$ in Eq. (A1), which can be maximized when some x_{jk} are zero. For this reason, the constraint $x_{jk} \geqslant 0$, $\forall j, k$, must be explicitly applied when solving this alternative formulation, and this leads to expressions in the solution that have "corners" or which are "nondifferentiable".

We do not wish here to solve this alternative formulation of DRAM, but only to investigate the number m of category-mode pairs $(j, k) \in \{1, 2, \ldots, J; 1, 2, \ldots, K\}$ such that $\hat{x}_{jk} > 0$. From Eq. (A4), this number satisfies

$$J \leqslant m \leqslant JK \tag{A5}$$

but stronger conditions on m can be found.

Using Lagrange multipliers λ_l, $l = 1, 2, \ldots, L$, to adjoin an equality resource constraint

$$R_l - \sum_j \sum_k x_{jk} y_{jkl} = 0 \qquad \forall l \tag{A6}$$

to function (A1), which is to be maximized, gives

$$H(x, y, z, \lambda) = \sum_j g_j(z_j) + \sum_j \sum_k \sum_l x_{jk} h_{jkl}(y_{jkl}) + \sum_l \lambda_l (R_l - \sum_j \sum_k x_{jk} y_{jkl})$$

Solutions for $\hat{y}(\lambda)$ satisfy

$$\frac{\partial H}{\partial y_{jkl}} = x_{jk} h'_{jkl}(y_{jkl}) - \lambda_l x_{jk} = 0$$

which gives

$$\hat{y}_{jkl} = h'^{-1}_{jkl}(\lambda_l) \qquad \forall j, k \text{ such that } x_{jk} > 0 \tag{A7}$$

Solutions for $\hat{x}(\lambda)$ are zero or satisfy

$$\frac{\partial H}{\partial x_{jk}} = g'_j(z_j) + \sum_l h_{jkl}(y_{jkl}) - \sum_l \lambda_l y_{jkl} = 0$$

which gives

$$\hat{z}_j = (g'_j)^{-1} \left\{ \sum_l [\lambda_l \hat{y}_{jkl} - h_{jkl}(\hat{y}_{jkl})] \right\}$$

Using Eqs. (A2), (A3), and (A7), this becomes

$$\hat{z}_j = Z_j(\mu_{jk})^{-1/(\alpha_j+1)} \qquad \forall j, k \text{ such that } x_{jk} > 0 \tag{A8}$$

where μ_{jk} is a function of λ similar to that defined by Eqs. (9) and (10). Because the left-hand side of Eq. (A8) is independent of k, it implies $(m - J)$ identities of the form $\mu_{jk_1} = \mu_{jk_2}$, in which there are only L unknowns λ_l, $l = 1, 2, \ldots, L$. In general therefore, solutions exist only if $(m - J) \leqslant L$. Combining this result with Eq. (A5) gives the following condition on the number of category-mode pairs that can be active:

$$J \leqslant m \leqslant \min(JK, J + L) \tag{A9}$$

For the data in Example 2 in Section 5, $J = 7$, $K = L = 2$, and inequality

(A9) is $7 \leqslant m \leqslant 9$. This implies that of the seven patient categories, not more than two can use more than one mode of care. For some definitions of categories and modes this result may be realistic, and we have made progress in solving models like this using nonsmooth optimization methods (Lemarechal and Mifflin 1978, Hughes *et al.* 1979). For Example 2, however, this result is unrealistic, because all categories of patients use both modes of care. Therefore, we have not pursued this formulation here.

B COMPUTER PROGRAMS AND SOLUTION EFFICIENCY

The procedures for model solution and parameter estimation described in Sections 3 and 4 have been implemented as computer programs. They are written in simple FORTRAN with many in-line comments, error handling, and full but suppressible diagnostic printout. They use no special software beyond simple matrix manipulation routines. Input and output files are read and written sequentially, and all files are formatted for easily understandable display. The programs are best used interactively, and a small utility program can quickly modify the input file when many model runs with different resource levels are required. Batch operation is equally possible.

Table B1 gives some statistics for the three principal programs, which solve the model with given parameters, estimate the level parameters X, Y, and estimate the power parameters α, β. They show that the average length of each routine is low (less than 60 statements) and that the fraction of comment code is high (more than 0.5). The total core load of each program is reasonable (less than 55K decimal bytes).

All three principal programs use an iterative solution and the running times therefore depend upon the starting values, the accuracy required of the solution, and the conditioning of the problem. For the model solution program, the running time additionally depends upon whether the dual constraint, Eq. (17) or (21), is applied and binding. Section 2.4 described how this constraint is handled computationally.

Table B2 gives typical running times for the three principal programs, used on problems of different sizes, when no diagnostic printout was requested and with arbitrary starting values (typically a first guess of $\lambda = 5$). Convergence is measured by the fractional change of

- The dual function $\hat{H}(\lambda)$, in the model solution program
- $\partial\hat{H}(\lambda)/\partial\lambda$, in the X, Y estimation program
- λ, in the α, β estimation program

and is usually fast. It is especially so for the model solution program (less than 15 CPU for a medium-sized problem), even when the solution lies on a dual constraint. No attempt has been made to speed up the parameter estimation programs, the second of which may converge slowly or not at all. But the fast

TABLE B1 Computer program statistics.

	Separate routines[a]	Number of FORTRAN statements[b]	Number of		Maximum numbers of				Total core load on PDP 11/70 (decimal bytes)
			Input files	Output files[c]	Patient categories	Treatment modes	Resource types	Data points	
Model solution	38	1,730 (1,061)	1	3	20	3	5		47,818
Estimation of X, Y	14	1,012 (557)	1	3	12	3	5	14	53,874
Estimation of α, β	16	1,241 (702)	1	3	12	3	5	14	44,264

[a]Including main line program, FUNCTION procedures, and matrix manipulation routines.
[b]The number of statements excluding COMMENTS is given in parentheses.
[c]One of these is suitable for display on a terminal during program execution.

TABLE B2 Typical running times of computer programs.

	Table showing run results	Dimensions of problem J	K	L	N	Precision of solution	Number of iterations to solution	CPU time to solution (sec)[a]
Model solution	3	6	1	1		10^{-5}	6	2.9
	4	6	1	1			4	2.7
	4	6	1	1			8	2.9
	7	7	2	2			8	12.8
	8	7	2	2			4	12.8
		5	1	2			4	5.9[b]
		7	2	2			6	14.1[b]
Estimating X, Y	3	6	1	1	14	10^{-3}	2[c]	14.7
	6	7	2	2	5		2[c]	18.7
Estimating α, β	3	6	1	1	14	$5 \cdot 10^{-2}$	528[d]	143.2
	6	7	2	2	10	10^{-3}	8	17.2

[a]With no diagnostic printout.
[b]In these runs the dual constraint was binding. In others it was not.
[c]Average number of iterations per data point.
[d]Very badly conditioned problem. Convergence is usually faster.

model solution program means that improved parameter estimation methods such as the one described in Section 4.5 are highly practical.

C FITTING FOUR PARAMETERS TO FOUR DATA POINTS

In this appendix we consider how to estimate the four model parameters X, Y, α, β from four data points (x_i, y_i), $i = 1, 2, 3, 4$, for the simplest possible DRAM when $J = K = L = 1$. This analysis extends and completes the discussion of Sections 3.2 and 3.3. We assume without loss of generality that $x_1 > x_2 > x_3 > x_4$ and $y_1 > y_2 > y_3 > y_4$. (If such an ordering of the data is not possible, it means that they do not satisfy the fundamental condition (26) on admissible solutions of DRAM.)

Equations (27) and (28) in Section 3.2 determine X, Y from x_1, x_2, y_1, y_2 when α, β are known. Substituting these results and the other two data points into Eq. (25), we get two nonlinear equations

$$\gamma_3(\alpha, \beta) = \gamma_4(\alpha, \beta) = 0 \tag{C1}$$

which determine α, β implicitly, where

$$\gamma_3(\alpha, \beta) = \left(\frac{y_1}{y_2}\right)^{\beta}\left[\left(\frac{x_1}{x_3}\right)^{\alpha+1} - 1\right] - \left(\frac{y_1}{y_3}\right)^{\beta}\left[\left(\frac{x_1}{x_2}\right)^{\alpha+1} - 1\right] - \left(\frac{x_1}{x_3}\right)^{\alpha+1} + \left(\frac{x_1}{x_2}\right)^{\alpha+1} \tag{C2}$$

$$\gamma_4(\alpha,\beta) = \left(\frac{y_1}{y_2}\right)^{\beta}\left[\left(\frac{x_1}{x_4}\right)^{\alpha+1} - 1\right] - \left(\frac{y_1}{y_4}\right)^{\beta}\left[\left(\frac{x_1}{x_2}\right)^{\alpha+1} - 1\right] - \left(\frac{x_1}{x_4}\right)^{\alpha+1} + \left(\frac{x_1}{x_2}\right)^{\alpha+1} \tag{C3}$$

Equations (C1) also define implicit functions $\beta_3(\alpha)$, $\beta_4(\alpha)$ which in turn define solutions $\tilde{\alpha}, \tilde{\beta}$ for α, β

$$\tilde{\beta} = \beta_3(\tilde{\alpha}) = \beta_4(\tilde{\alpha}) \tag{C4}$$

For successful solution of Eq. (C4), two sets of existence conditions need to be established. First, we must find conditions for Eqs. (C1) to have a solution $\beta > 0$, assuming the existence of a solution $\alpha > 0$. Second, we must find conditions for Eq. (C4) to have a solution $\alpha > 0$. When the second condition is satisfied, the first condition will ensure $\beta > 0$.

The first conditions follow from inspecting the derivatives $\partial\gamma_3/\partial\alpha$, etc. We find that sufficient (but more than necessary) conditions for $\beta > 0$ given $\alpha > 0$ are

$$\left(\frac{(x_1/x_3) - 1}{(x_1/x_2) - 1}\right) > \left(\frac{\ln(y_1/y_3)}{\ln(y_1/y_2)}\right) \tag{C5}$$

$$\left(\frac{(x_1/x_4) - 1}{(x_1/x_2) - 1}\right) > \left(\frac{\ln(y_1/y_4)}{\ln(y_1/y_2)}\right) \tag{C6}$$

The second conditions follow from lower, upper, and asymptotic estimates of the functions $\beta_3(\alpha)$, $\beta_4(\alpha)$.

$$\beta_j^{\min}(\alpha) = \left\{\ln\left[\frac{(x_1/x_j)^{\alpha+1} - 1}{(x_1/x_2)^{\alpha+1} - 1}\right] + \ln\left[\frac{\ln(y_1/y_j)}{\ln(y_1/y_2)}\right]\right\} \Big/ \left(\ln\frac{y_2}{y_j}\right) \tag{C7}$$

$$\beta_j^{\max}(\alpha) = \left\{\ln\left[\frac{(x_1/x_j)^{\alpha+1} - 1}{(x_1/x_2)^{\alpha+1} - 1}\right]\right\} \Big/ \left(\ln\frac{y_2}{y_j}\right) \tag{C8}$$

$$\beta_j^{\alpha\to\infty}(\alpha) = (\alpha + 1)\frac{\ln(x_2/x_j)}{\ln(y_2/y_j)} \qquad j = 3, 4 \tag{C9}$$

If $\beta_3^{\infty}(\alpha) > \beta_4^{\infty}(\alpha)$, then $\beta_4(0) > \beta_3(0)$ will guarantee a solution $\tilde{\alpha} > 0$ to Eq. (C4). This condition is depicted in Figure C1. Conversely, if $\beta_4^{\infty}(\alpha) > \beta_3^{\infty}(\alpha)$, then $\beta_3(0) > \beta_4(0)$ will guarantee a solution. Both of these sufficient and necessary conditions (which must be computed numerically) can be approximated by sufficient but more restrictive conditions (which need not be computed numerically)

$$\beta_4^{\min}(0) > \beta_3^{\max}(0) \qquad \text{or} \qquad \beta_3^{\min}(0) > \beta_4^{\max}(0)$$

In order to illustrate the approach, we consider the data shown in Table C1 which satisfy conditions (26), (C5), and (C6). In addition, $\beta_3^{\infty}(\alpha) > \beta_4^{\infty}(\alpha)$ and $\beta_4(0) > \beta_3(0)$, thereby guaranteeing solutions $\tilde{\alpha}, \tilde{\beta} > 0$. On Figure C1 are plotted values of $\beta_3(\alpha)$, $\beta_4(\alpha)$ obtained by solving Eqs. (C1) by the following iteration

$$\beta_j(i+1) = 2\bar{\beta}_j(i+1) - \beta_j(i) \qquad j = 3, 4, \quad i = 1, 2, \ldots$$

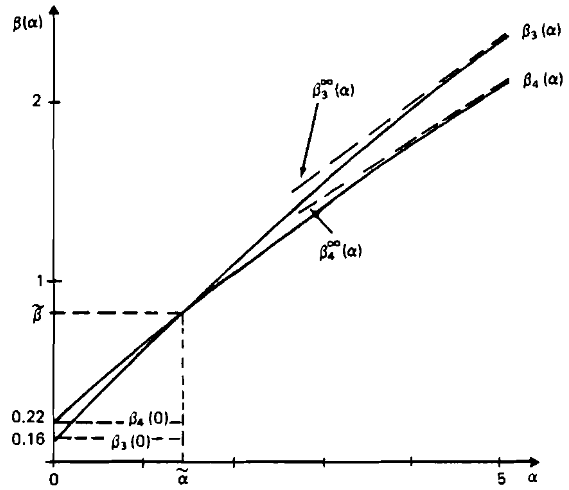


FIGURE C1 Solution of Eq. (C4) for data from Table C1.

where

$$\bar{\beta}_j(i+1) = \frac{\ln\left[1 + \frac{(x_1/x_j)^{\alpha+1} - (x_1/x_j)}{(x_1/x_2)^{\alpha+1} - 1} - \frac{(x_1/x_j)^{\alpha+1} - (x_1/x_2)^{\alpha+1}}{(x_1/x_2)^{\alpha+1} - 1}(y_2/y_1)^{\beta_j(i)}\right]}{\ln(y_2/y_j)}$$

The solution to Eq. (C4) is found at $\tilde{\alpha} = 1.60$, $\tilde{\beta} = 0.83$, $X = 1.05$, and $Y = 1.08$, although it is not very accurately determined because the problem is rather ill-conditioned. This is seen in the approximate equality of $\beta_3^\infty(\alpha) = 0.386(\alpha + 1)$ and $\beta_4^\infty(\alpha) = 0.355(\alpha + 1)$, and in the very flat intersection in Figure C1. Nevertheless, the estimated values are close to the true parameter values (shown in Table C1) used to derive the four data points.

TABLE C1 Data for test of parameter fitting. Solutions of the simplest possible DRAM with $x = 1, y = 1, \alpha = 2, \beta = 1$.

	$i = 1$	$i = 2$	$i = 3$	$i = 4$
x_i	0.90	0.70	0.40	0.20
y_i	0.84	0.51	0.12	0.015
$R_i = x_i y_i$	0.756	0.357	0.048	0.003

D UNBIASED REGRESSION ESTIMATORS

In the estimation of power parameters α, β in Section 4.3, we assumed that α, β are constant across the areas of a region, and then we performed regression analysis on the cross-sectional data. However, even if this assumption is incorrect and α, β are different in different areas, we can show that this procedure still yields useful regional estimates.

We define the indices $j = 1, 2, \ldots, J$ areas or subregions, and $i = 1, 2, \ldots, N$ observations in each area, and suppose that data $x_j(i), y_j(i)$ satisfy the linear model

$$y_j(i) = b_j x_j(i) + \epsilon_j(i) \tag{D1}$$

in which $\epsilon_j(i)$ are uncorrelated random disturbances with zero mean and variance σ^2. The unknown parameter b_j is different for different areas. Nevertheless, we assume that it is constant and we form the usual least-squares estimate

$$\hat{b} = \left(\sum_j X_j'X_j\right)^{-1} \sum_j X_j'Y_j \tag{D2}$$

in which $X_j = \{x_j(1), \ldots, x_j(N)\}'$ and $Y_j = \{y_j(1), \ldots, y_j(N)\}'$. We now investigate the properties of $\hat{b}$ when the unknown parameters b_j are actually random samples from a normal or Gaussian probability density function with mean m and variance v^2:

$$b_j \sim N(m, v^2)$$

Combining Eqs. (D1) and (D2) gives

$$(\hat{b} - m) = \left(\sum_j X_j'X_j\right)^{-1} \left\{\sum_j X_j'X_j(b_j - m) + \sum_j X_j'E_j\sigma\right\}$$

whence the result: EXPECTATION $(\hat{b} - m) = 0$; the estimator b is an unbiased estimator of the mean regional parameter m.

Equations (55) and (56) in Section 4.3 are like Eq. (D1). The functions corresponding to b_j are $(\alpha_j + 1)^{-1}$ and $(\beta_{jkm} + 1)^{-1}$ which are estimated without bias, subject to the above assumptions. Additionally, we may show that

$$\text{EXPECTATION } (\hat{b} - m)^2 \sim \frac{\sigma^2}{JN} + \frac{v^2}{J}$$

The first term on the right-hand side is the usual residual variance term, and the second arises from the uncertainty about b_j.

E SENSITIVITY OF THE SOLUTION TO PARAMETER CHANGES

The parameter estimation procedure described in Section 4.5 needs expressions for the sensitivities of the solutions $\hat{x}, \hat{y}$ to a change in a parameter $p \in P = \{X, Y, \alpha, \beta\}$. These expressions are derived below.

The total sensitivity derivatives can be written as the sum of two sets of partial derivatives

$$\frac{d\hat{x}_{jk}}{dp} = \frac{\partial \hat{x}_{jk}}{\partial p} + \sum_m \frac{\partial \hat{x}_{jk}}{\partial \hat{\lambda}_m} \frac{\partial \hat{\lambda}_m}{\partial p}$$

$$\frac{d\hat{y}_{jkl}}{dp} = \frac{\partial \hat{y}_{jkl}}{\partial p} + \sum_m \frac{\partial \hat{y}_{jkl}}{\partial \hat{\lambda}_m} \frac{\partial \hat{\lambda}_m}{\partial p}$$

The first term in each equation is the partial derivative when the Lagrange

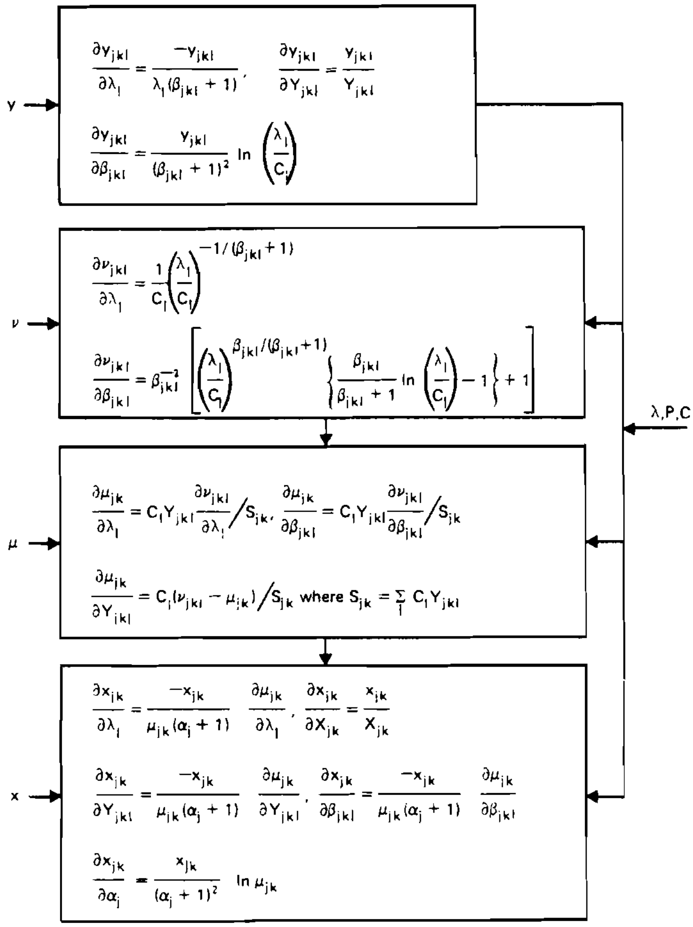


FIGURE E1 Calculation of sensitivity derivatives (superscript carets (ˆ) are omitted for clarity).

multipliers λ are held constant. The second term in each equation reflects the sensitivity of the solution to changes in the Lagrange multipliers.

In order to obtain the terms $\partial\hat{\lambda}/\partial p$ we note that, at the solution point $\lambda = \hat{\lambda}$,

$$\frac{\partial \hat{H}(\lambda)}{\partial \hat{\lambda}_l} = F_l[\hat{x}(\hat{\lambda}), \hat{y}(\hat{\lambda})]$$

$$= R_l - \sum_j\sum_k \hat{x}_{jk}\hat{y}_{jkl} = 0 \qquad \forall l$$

Differentiating this result with respect to $p \in P$ gives

$$\frac{\mathrm{d}}{\mathrm{d}p}\left[\frac{\partial \hat{H}(\lambda)}{\partial \hat{\lambda}_l}\right] = \frac{\partial}{\partial p}\left[\frac{\partial \hat{H}(\lambda)}{\partial \hat{\lambda}_l}\right] + \sum_m \frac{\partial^2 \hat{H}(\lambda)}{\partial \hat{\lambda}_l \partial \hat{\lambda}_m}\frac{\partial \hat{\lambda}_m}{\partial p} = 0 \qquad \forall l$$

whence $(\partial\hat{\lambda}/\partial p) = -(\hat{H}_{\lambda\lambda})^{-1}(\partial\hat{H}_{\lambda}/\partial p)$ or

$$\frac{\partial\hat{\lambda}_m}{\partial p} = \sum_l (\hat{H}_{\lambda\lambda})^{-1} \sum_j \sum_k \hat{x}_{jk} \frac{\partial\hat{y}_{jkl}}{\partial p} + \hat{y}_{jkl} \frac{\partial\hat{x}_{jk}}{\partial p}$$

$(\hat{H}_{\lambda\lambda})^{-1}$ is the inverse Hessian matrix which is calculated during the solving of the model. The other terms are simply the other group of partial derivatives that follow straightforwardly from Eqs. (7)–(10). The only difficulty is in organizing the computation in the most convenient way. Figure E1 depicts a possible scheme.

The calculations are considerably simplified by the presence of many zero terms. Most obviously,

$$\frac{\partial y_{jkl}}{\partial X_{jk}} = \frac{\partial y_{jkl}}{\partial \alpha_j} = 0$$

$$\frac{\partial \nu_{jkl}}{\partial X_{jk}} = \frac{\partial \nu_{jkl}}{\partial Y_{jkl}} = \frac{\partial \nu_{jkl}}{\partial \alpha_j} = 0$$

$$\frac{\partial \mu_{jk}}{\partial X_{jk}} = \frac{\partial \mu_{jk}}{\partial \alpha_j} = 0 \qquad \forall j, k, l$$

Less obviously

$$\frac{\partial y_{jkl}}{\partial Y_{\bar{j}\bar{k}\bar{l}}} = \frac{\partial y_{jkl}}{\partial \beta_{\bar{j}\bar{k}\bar{l}}} = 0$$

$$\frac{\partial \nu_{jkl}}{\partial \lambda_{\bar{l}}} = \frac{\partial \nu_{jkl}}{\partial \beta_{\bar{j}\bar{k}\bar{l}}} = 0$$

$$\frac{\partial \mu_{jk}}{\partial \beta_{\bar{j}\bar{k}\bar{l}}} = \frac{\partial \mu_{jk}}{\partial Y_{\bar{j}\bar{k}\bar{l}}} = 0$$

$$\frac{\partial x_{jk}}{\partial X_{\bar{j}\bar{k}}} = \frac{\partial x_{jk}}{\partial Y_{\bar{j}\bar{k}\bar{l}}} = \frac{\partial x_{jk}}{\partial \alpha_{\bar{j}}} = \frac{\partial x_{jk}}{\partial \beta_{\bar{j}\bar{k}\bar{l}}} = 0$$

for $(j, k, l) \neq (\bar{j}, \bar{k}, \bar{l})$. Unfortunately, the matrices of total derivatives

$$\left\{\frac{\mathrm{d}\hat{x}_{jk}}{\mathrm{d}p}\right\}, \left\{\frac{\mathrm{d}\hat{y}_{jkl}}{\mathrm{d}p}\right\}$$

have in general no zero terms because of the dependence of each Lagrange multiplier upon every parameter. Together, they have $(JK + JKL)(2JKL + JK + J)$ terms, but Section 4.5 shows that not all these elements need to be stored.

F LIST OF PRINCIPAL SYMBOLS

Symbols used only in the appendixes are not included here.

Symbol	Definition	Page of first appearance
i =	1, 2, . . . , N, iterations, times, regions, data points	9, 21, 21, 24
j =	1, 2, . . . , J, patient categories	3
k =	1, 2, . . . , K, modes of care	3
l =	1, 2, . . . , L, resource types	3
A, B	decomposition of $\hat{H}_{\lambda\lambda}$	6
a_{lm}, b_{lm}	elements of A, B	6
A_{jkl}, B_{ml}	expressions relating α to γ and β to η	24
a^x_{jk}, a^y_{jkm}	constant terms in regression estimation of γ, η	24
C_l	marginal unit cost of type l resource, when all needs are met	4
d^i	Newton direction at iteration i	9
$F_l(x, y)$	function in constraint equation	4
$f_l(\lambda)$	function specifying θ_l	21
$g_{jk}(x)$	functions measuring the benefits of increasing service levels	4
$h_{jkl}(y)$	functions measuring the benefits of increasing supply levels	4
$H(x, y, \lambda)$	Lagrangian function	5
$\hat{H}(\lambda)$	Lagrangian when $x = \hat{x}(\lambda)$, $y = \hat{y}(\lambda)$	5
$\hat{H}_\lambda, \hat{H}_{\lambda\lambda}$	gradient and Hessian of $H(\lambda)$ when $x = \hat{x}(\lambda)$, $y = \hat{y}(\lambda)$	5
$\bar{H}_{ml}$	ml-th element of inverted Hessian matrix	23
H_{zz}	matrix of second derivatives of Lagrangian with respect to primal variables	6
$J(P)$	function of squared prediction errors	27
$\bar{L}$	set of active resource constraints	9
M	finance for purchasing resources	7
$P(i)$	population in region i	21
P	parameter set $\{X, Y, \alpha, \beta\}$	26
p, q =	$(x/X), (y/Y)$	12
R_l	available resource of type l	4
r_l	excess resource of type l	7
S	statistics of ξ_1, ξ_2 processes	27
s	convergence coefficient	10
t	step-size coefficient	9
$U(x, y)$, $U(x, y, P, R, C)$	function which is maximized by DRAM	4, 26
v_m	sensitivity derivative vectors	28
$X_{jk}, x_{jk}, \hat{x}_{jk}$	ideal, actual, optimal service levels	4, 5
$Y_{jkl}, y_{jkl}, \hat{y}_{jkl}$	ideal, actual, optimal supply levels	4, 5
$\bar{X}, \bar{Y}$	regional ideal levels	21

Z_j	ideal service levels summed across modes	21
z	arbitrary positive vectors	6
α_j, β_{jkl}	model power parameters	4
$\gamma_{jkl}, \eta_{jkml}$	empirical elasticities	23
δ_{lm}	Kronecker delta function	6
$\epsilon^x_{jk}, \epsilon^y_{jkm}$	error terms in regression estimation of γ, η	24
θ_l, ϕ_l	additional information for estimating X, Y	20, 22
$\tan\Theta$	$= (\beta + 1)/(\alpha + \beta + 2)$	12
$\tan\Phi$	$= (\alpha + 1)/(\alpha + \beta + 2)$	12
$\tan\Psi$	$= (\alpha + 1)/(\beta + 1)$	13
ω	$= \ln p_2/\ln p_1$	15
τ	$= (1 - \omega)\ln q_1$	15
μ_{jk}, ν_{jkl}	functions in expression for $\hat{y}$	5
$\lambda_l, \hat{\lambda}_l$	actual and optimal Lagrange multipliers	5
$\tilde{\lambda}_l$	minimum Lagrange multipliers	22
$\zeta(\beta, q)$	$= \ln\left[\left(1 + \frac{1}{\beta}\right)q^{-\beta} - \frac{1}{\beta}\right]$	15
ξ_1, ξ_2	random processes perturbing x, y	27
$\rho^x_{ijk}, \rho^y_{ijkl}$	weighting terms in $J(P)$	27

ENERGY AND ENTROPY FLUXES IN COAL GASIFICATION AND LIQUEFACTION PROCESSES

Hans Voigt

PREFACE

In the long-term studies on energy systems performed at IIASA, scenarios that provide for substitutes for fossil oil and gas are considered. In the future coal is expected to contribute to energy supplies to a greatly increasing extent only if it is converted to liquid or gaseous fuels or electricity. Coal conversion systems are rather complex, not only internally but also with respect to their exchanges with the environment; some use auxiliary energy, others yield by-products. Therefore, the evaluation of such systems is not a simple task and the comparison of very different systems – different in the nature of inputs and outputs – must not be reduced to a comparison of energy efficiencies.

Moreover, because these studies cover a long time period, it is necessary to estimate the potential development of related processes in order to determine the inputs required for producing substitute fuels. There are physical and chemical limitations to potential improvement. This paper outlines these constraints and provides means for the evaluation and comparison of different fuel synthesis processes, especially regarding methanol. The possibility of adding energy from nuclear or solar primary energy sources to such processes is discussed and the advantages are assessed.

BASIC ANALYSIS

Coal, being the largest fossil energy resource, plays an important role in all future energy supply scenarios. In a solid state it cannot be used to a greatly increasing extent as a fuel for the final consumer. If converted to liquid and gaseous fuels or to electricity it is more suitable. Electricity generation from coal is very important in this context; however, this is the state of the art and it is therefore not considered in more detail in this paper. In generating electricity from coal there are constraints, for economic reasons at least, resulting from the location of coal resources and from the relatively high transportation costs for coal and electricity. Liquid and gaseous fuels produced from coal, however, could well serve as substitutes for fossil oil and gas when the latter fuels become scarce. The substitute fuels could be produced almost free of sulfur.

Gasification and liquefaction of coal have already been carried out. The principal processes used commercially are those of Lurgi, Winkler, and Koppers-Totzek for gasification, and that of Fischer-Tropsch for liquefaction. These and similar processes are being developed to improve their economy and efficiency. Furthermore, the possibility of adding external energy from nuclear reactors or solar collectors to such processes is being investigated. The advantages of the latter procedure over the autothermal coal conversion procedure (i.e., no energy other than that of coal is supplied to the process) should be greater fuel yields from a given amount of coal, decreased carbon dioxide emissions, and also possibly certain economic benefits.

Scenarios of world energy supplies in, for example, 50 years, take into account that several terawatts (TW) of methanol will have to be produced from coal and nuclear or solar energy (Haefele and Sassin 1977); it is, therefore, essential to search for efficient and economic processes for methanol production. In this study, the natural limits of these processes are evaluated against a background of the relevant thermodynamic and chemical laws. This allows a judgment to be made about the "quality" of a process and the limits to its further development. Several proposed processes, especially molten-iron bath gasification (being developed by Humboldt-Wedag in the FRG), are examined, particularly in relation to the coal and additional energy they require.

The processes for fuel production from coal have to be considered, among others, from three specific aspects: energetic, exergetic, and chemical. It is energy that is usually considered in the evaluation of fuel production processes. The energy efficiency η (i.e., the energetic value of the yield over the energetic value of the expense) is used to characterize a process. The energy efficiencies of current autothermal gasification and liquefaction processes range from approximately 0.4 to 0.75 (i.e., this fraction of the chemical energy of coal is to be found in the products, or the energy expense – coal – is 1.3 to 2.5 times greater than the energy yield).

In general, energy efficiency $\eta = 1$ is not the natural limit. Therefore, it is not sufficient to estimate the potential improvement of a process on the basis of its energy efficiency alone (Voigt 1978). Rather, such an estimate has to be made by taking into account the entropy flows that a system exchanges with its environment. This enables the entropy production, which is the absolute measure of the system's thermodynamic quality, to be calculated. Certainly, entropy is not as easily visualized in combination with energy and is not as easily handled by non-specialists. Therefore, quantitative considerations are given preferably in terms of exergy, which is defined as

$$\text{Exergy} = E - T_0 S \qquad (1)$$

Exergy can be interpreted as the maximum work that can be provided by energy E that is accompanied by entropy S, if it is possible to exchange heat with an environment of temperature T_0. Exergy has the same dimension and order of magnitude as energy, it is a measure of the "quality" of energy. The ratio of exergy yield to exergy expense of a system or process is called reversibility ϵ (or exergy efficiency or second law efficiency); ϵ represents the proximity of a process to the thermodynamic limit: $\epsilon = 1$ for an ideal, reversible process; $\epsilon < 1$ for a real, irreversible process. The degree of reversibility indicates the potential for improvement of a system. The formalism for this evaluation is well known in technical thermodynamics and a single general description is given in Voigt (1978). For current autothermal gasification and liquefaction processes, the reversibility ranges between 0.35 and 0.65 (i.e., this fraction of the exergy of the coal used is found in the gas and the liquid products).

From the chemical aspect, the number of carbon atoms that are contained in a fuel are taken into account. Fossil coal can be characterized approximately in relation to its energetically relevant constituents, by the formula CH_y, with y ranging between 0.5 and 1. Between 80 and 90 percent of the exergy of coal can be attributed to carbon. Methane, methanol, and gasoline contain 0.48, 0.56, and 0.60 units of carbon, respectively (in terms of the exergy of the oxidation of carbon to carbon dioxide), in 1 exergy unit of fuel (see Figure 1). Hydrogen, of course, contains no carbon. Carbon monoxide, which is not so

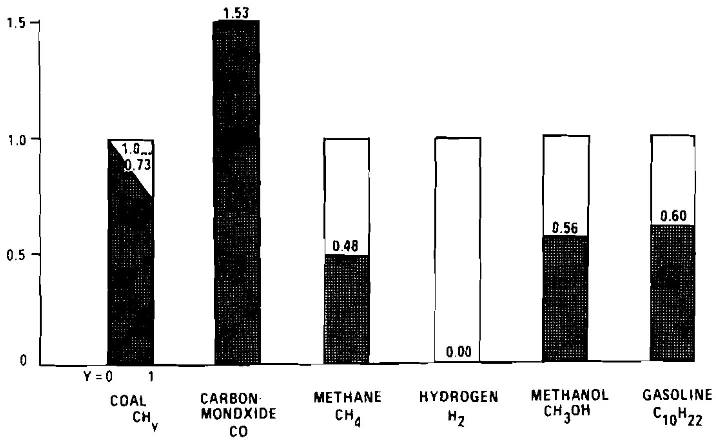


FIGURE 1 Carbon content (shaded area) of different fuels in terms of exergy.

important as fuel but essential for methanol synthesis, has less exergy than the contained carbon had in its elementary form, since one-third of the exergy has been released at the stage of carbon conversion to carbon monoxide. From Figure 1 we can see that 1.53 exergy units of carbon are "contained" in (i.e., required for) 1 exergy unit of carbon monoxide.

The figures given above represent the minimum amount of carbon necessary for the synthesis of those fuels (i.e., required for stoichiometric processes with no carbon losses). Technical processes have carbon losses, mostly in the form of carbon dioxide. The blank areas in Figure 1 indicate the minimum amount of exergy (i.e., for an ideal, reversible process) that has to be added from other sources if only the minimum carbon demand were expended. Real processes are irreversible and require a larger amount of exergy than reversible processes. Thus, the data in Figure 1 can serve as a standard against which real processes may be measured. These data may also be regarded as the asymptotic limits to further, long-term development of processes. So, it appears that the liquid fuels methanol and gasoline do not significantly differ from each other from the standpoint of carbon demand and the exergy that can be added. However, these liquid fuels differ for technical, economic, and environmental reasons and also in relation to their penetration of the market.

To evaluate real or conceivable processes, the appropriate reversible process should be used as a yardstick to measure the amount of exergy required (Voigt 1978). Consider a general fuel conversion system (see Figure 2) that is fed with coal and heat of temperature T_1 (expense), that produces a fuel (yield) – methanol in this case – and in which all other exchanges with the environment are counted as waste (dissipation). Each of these three streams (expense, yield, dissipation) is characterized by energy E, entropy S, and the number of carbon atoms N_C it contains. The conservation laws of thermodynamics and chemistry should then be applied to the processes. For

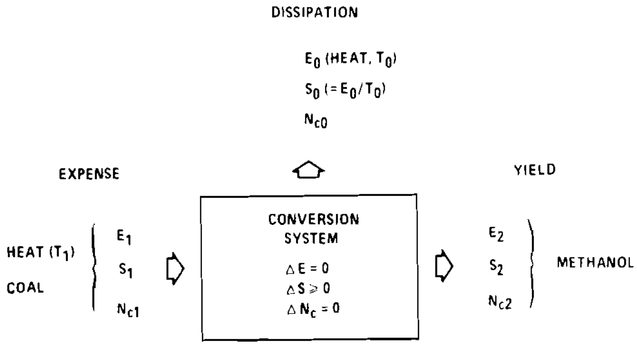


FIGURE 2 Energy (E), entropy (S), and carbon atoms (N_C) that a methanol production system exchanges with its environment.

stationary processes (all variables constant in time), using the notation given in Figure 2, these laws read:

$$E_1 = E_0 + E_2 \tag{2}$$

$$S_1 + \Delta S = S_0 + S_2 \quad \Delta S \geqslant 0 \tag{3}$$

$$N_{C1} = N_{C0} + N_{C2} \tag{4}$$

ΔS is the entropy production of the system and is not negative for the second law of thermodynamics; since energy and carbon atoms are neither produced nor annihilated, $\Delta E = 0, \Delta N_C = 0$. We assume that $N_{C0} = 0$, that no carbon atoms are wasted, and that $E_0 = T_0 S_0$, all wasted energy is heat of environmental temperature T_0. Then, taking into account the thermodynamic properties of carbon and methanol, we arrive at a relation between E_1 and E_2 that depends on T_1 (temperature of expended heat) and on ΔS (entropy production of the system). In Figure 3, the energy expense E_1 is plotted (left-hand scale) against the temperature T_1 and normalized for the yield of 1 energy unit of methanol, $E_2 = 1$. Of the total energy expense, 0.54 units are expended as coal (if it were devoid of hydrogen), the remainder is heat. The curve $\epsilon = 1$ is valid for reversible processes, $\Delta S = 0$. For example, if heat of 800 Kelvin (K) is available, 0.68 units of heat have to be added to the 0.54 energy units of carbon, resulting in a total energy expense of 1.22 units for 1 energy unit of methanol; therefore, 0.22 units of energy are inevitably wasted. This is the absolute minimum dissipation of energy and serves as the yardstick for real, irreversible processes. The corresponding energy efficiency η (0.82 in this case) can be read from the right-hand scale.

For a lower degree of reversibility, for example, $\epsilon = 0.5$, if the coal expense is held at the chemical minimum (0.54), the expense of heat required is more than doubled and increases to 2.2 units of 800 K heat. So, the total

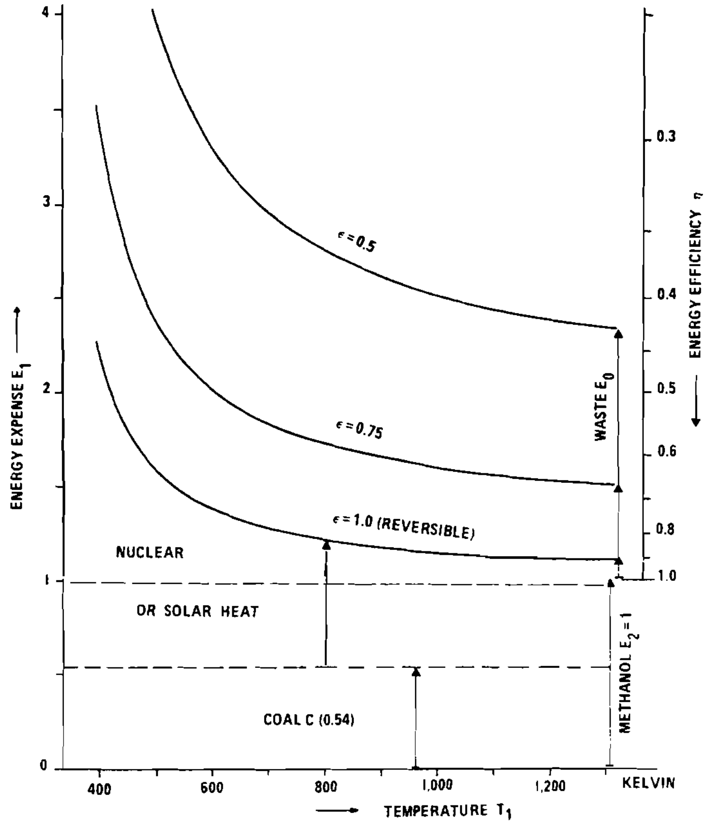


FIGURE 3 Energy expense E_1 (heat + coal) required to produce 1 unit of methanol for reversible ($\epsilon = 1$) and irreversible ($\epsilon < 1$) processes.

energy expense is 2.8 units for 1 unit of methanol, 1.8 units being wasted (see curve $\epsilon = 0.5$ in Figure 3).

THE REFERENCE CASE

Figure 3 illustrates how real or proposed methanol production processes with a known coal and heat input can be evaluated to determine how "good" the processes are (i.e., how far they are from natural limits). To proceed further, we take into account some more practical conditions, and, by making plausible assumptions about the main subsystems, arrive at an estimate of the energy efficiency that could be attained in the future. This is dependent on several factors.

Given certain technologies, in many cases reversibility can be improved by extending the equipment (e.g., enlarging the heat transfer area or using an expansion turbine instead of a throttle valve), which usually implies increasing capital investment. Therefore, the design of capital-intensive thermodynamic

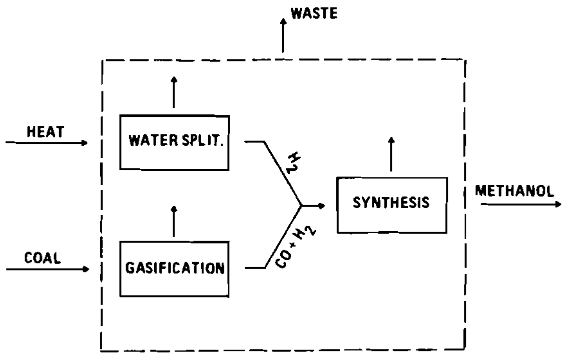


FIGURE 4 Main subsystems for methanol production from heat and coal.

equipment (e.g., thermal power plants) is made after carefully balancing the capital cost against the resulting increase in the product's value. A formalism for such an optimization procedure is given in El-Sayed and Evans (1968). With a long-term perspective, however, technological conditions cannot be considered as fixed. Through research and development, new ideas, new processes, and new materials are produced, all of which increase the efficiency and simultaneously decrease the extension and cost of equipment. The evolution of steam engines (both piston-engines and turbines) provides a good example of this. Therefore, for our estimate of reversibility, the basic thermodynamic and chemical principles, but not the technological or economic conditions, are regarded as fixed. As a consequence, the subsystems of the fuel conversion processes considered are characterized primarily by their task or function, rather than by fixed techniques.

In present coal gasification plants, hydrogen requirements are covered by carbon monoxide shifting. Since this is coupled to carbon dioxide production (i.e., wastage of carbon atoms), which should eventually be avoided, additional hydrogen production that is independent of carbon has to be provided. Therefore, the main subsystems for methanol production considered are gasification (including carbon monoxide shifting if it exists), water splitting, and synthesis (see Figure 4).

To achieve ideal conditions for gasification processes – no wastage of carbon, $N_{C_0} = 0$, and, simultaneously, no entropy production, $\Delta S = 0$ – it would be necessary to take up entropy (together with heat) from the environment, $E_0 < 0$, therefore, $\eta > 1$. This is because of the entropy balance in which the entropy of one mole of the products is larger than that of the inputs. It is unlikely that this will become technically feasible, because a type of reversible heat pump would have to be included in the system. Therefore, instead of taking $\Delta S = 0$ for the reference case, we prefer to take $E_0 = 0$, where no energy is wasted (i.e., $\eta = 1$). Thus, the reversibility is approximately 0.9, which is still a satisfactory figure. Besides carbon, which has already been

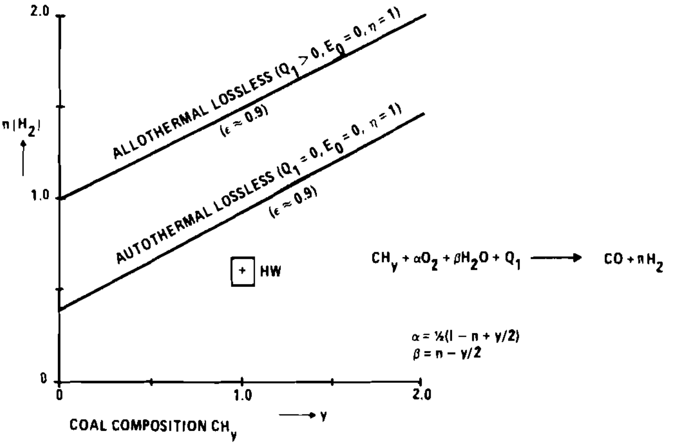


FIGURE 5 Hydrogen content n of synthesis gas, $CO + nH_2$, as a function of coal composition CH_y, for lossless ($E_0 = 0$, $\eta = 1$) autothermal ($Q_1 = 0$) and allothermal ($Q_1 > 0$) gasification (HW: Humboldt-Wedag gasification).

considered, the only other element of coal having major energetic and exergetic relevance is hydrogen. The hydrogen content ranges from 0.5 to 1 atom of hydrogen per atom of carbon; brown coal and lignites have on average 0.95 and pit coal and anthracites have 0.7 (Nesterov and Salmanov 1977). The essential feature of coal is that, relative to the energy of oxidation, hydrogen is bound very loosely to the carbon. In gasification processes producing synthesis gas for methanol synthesis, it is desirable to obtain a gas with a large hydrogen content, since this provides hydrogen that would otherwise have to be generated in other ways.

In Figure 5 the composition of synthesis gas, $CO + nH_2$, for two types of gasification processes is plotted against the hydrogen content of the coal used, y, according to the formula CH_y. The lower line represents autothermal processes (i.e., no energy other than that of the coal is supplied to the gasification process, $Q_1 = 0$) that are "lossless," $E_0 = 0$ (i.e., no energy is dissipated). The energy efficiency for the gasification subsystem, therefore, is $\eta = 1$, and the reversibility is $\epsilon \approx 0.9$. The hydrogen content of the product gas ranges from $n = 0.4$ molecules for pure carbon input to $n = 0.95$ molecules for coal input of composition CH. Allothermal processes (i.e., extraneous heat is added to the process, $Q_1 > 0$, see Figure 5, upper line) permit larger amounts of water to be added. If carried out without energy losses, $E_0 = 0, \eta = 1$, these processes yield a maximum of $n = 1$ to 1.5 hydrogen molecules for coal of composition C and CH ($y = 0$ and $y = 1$), respectively. For methanol synthesis, the hydrogen demand is $n = 2$ molecules of hydrogen; therefore, if coal with a large hydrogen content is used, only one-half of a hydrogen molecule has to be provided from other sources. Figure 5 extends to $y = 2$ (i.e., CH_2 as source composition). CH_2 no longer represents coal but mineral oil, and corresponds approximately to the present method of methanol production. However, the use of oil is exactly what should be avoided in the future. The composition of the gas for the molten-iron bath gasification process, to be dealt with later, is indicated by the cross (HW) in Figure 5.

For the water-splitting subsystems (see Figure 4), if electricity is expended, the technically attainable reversibility is estimated to be approximately $\epsilon = 0.75$, which corresponds to an energy efficiency of $\eta = 0.9$ (Getoff 1977). The conversion of heat to electricity in large thermal power plants is carried out today with an energy efficiency of $\eta = 0.40$ for $T_1 = 800\,K$ and $\eta = 0.32$ for $T_1 = 600\,K$; this corresponds to a reversibility of $\epsilon = 0.64$ in both cases. Although improvements in thermal power plants are also to be expected in the future, for the moment we shall retain these figures; the influence of an improvement is discussed later. Therefore, for the total water-splitting subsystem, starting with heat, we take an overall reversibility of $\epsilon = 0.5$ as the reference case. This could also be valid for thermochemical water-splitting processes developed in the future.

The synthesis of methanol from synthesis gas represents the state of the art. We take as the reference case a situation where no matter is lost, where the energy and the exergy differences between the (cold) synthesis gas and the liquid methanol are lost but no auxiliary energy is supplied. This gives a reversibility of $\epsilon = 0.96$, which is a very satisfactory figure, and an energy efficiency of $\eta = 0.85$.

For the reference case (see Figure 6), the total energy expense (upper line) and the shares of coal and heat are plotted against the hydrogen content y of the coal used. These lines are valid for lossless autothermal gasification, given heat of temperature 800 K. The importance of hydrogen in coal becomes obvious from a glance at Figure 6. For the case of coal of composition C (e.g., coke), there must be an expense of 0.54 energy units of this coal plus 1.75 units of heat, making a total of 2.3 units for the production of one energy unit of methanol or an energy efficiency of $\eta = 0.43$. In this case, the reversibility, given in Figure 3, is $\epsilon \approx 0.6$. For hydrogen-rich coal, e.g., coal of composition CH, 0.75 energy units of this coal plus 1.2 units of heat are necessary, giving an overall efficiency of $\eta = 0.52$ and a reversibility of $\epsilon = 0.66$.

Figure 6 also includes figures related to the molten-iron bath gasification process (from a private communication with R. Pfeiffer, KHD Industrieanlagen AG, Humboldt-Wedag). In this process, which is similar to steel-making processes, oxygen and steam are blown into a bath of molten iron and dissolved carbon at a temperature of approximately 1,600 K. Under such conditions, the gases react with the carbon to form carbon monoxide and hydrogen, and the generation of carbon dioxide can be avoided. The carbon extracted during the bath process is replaced continuously by granulated coal, which is also blown into the bath. All types of coal are considered to be suitable. During the process, the sulfur content of the coal combines with and is thus removed with the slag, and one can expect that almost

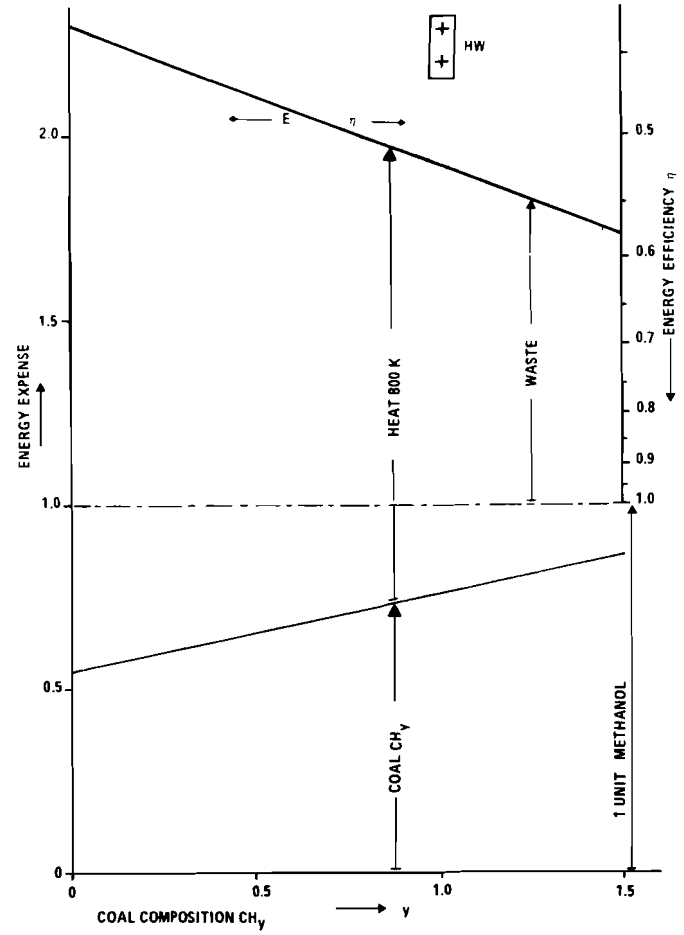


FIGURE 6 Energy expense (800 K heat + coal) required to produce 1 unit of methanol, as a function of coal composition CH_y, for lossless autothermal (upper line) and Humboldt-Wedag gasification (HW).

no carbon will be lost. The energy lost as heat from the bath is small compared to the large energetic throughput (about 10^7 W/m^2 of molten-iron bath) which is 30 times the black radiation at 1,600 K. The only difficulty, with respect to energy, is that the product gases (and slag) are emitted at that high temperature, taking with them about 12 percent of the energetic throughput as sensible heat. If, under ideal conditions, all this sensible heat could be fed back to the process (for preheating the input), we would arrive at the lossless autothermal process already considered (upper line in Figure 6, lower line in Figure 5). If, however, under the worst conditions, all the sensible heat is dissipated, this energy must be provided by the gasification reaction, which then has to be made exo-

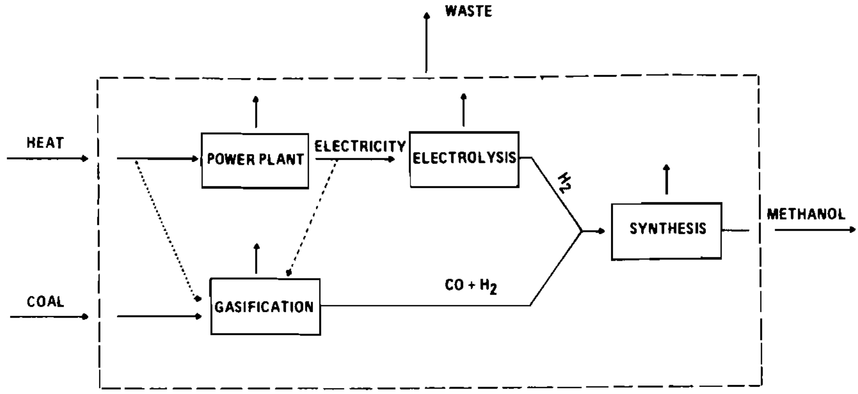


FIGURE 7 Configurations for methanol production using allothermal gasification.

thermally. As a consequence, less steam can be applied and hence less hydrogen can be produced (indicated by the cross (HW) in Figure 5). Since the hydrogen production is lower, extra hydrogen must be produced in other ways and thus additional auxiliary energy – heat of temperature 800 K – is required (indicated by the upper cross (HW) in Figure 6). As a compromise, one could consider transforming the sensible heat of the effluent gases into electricity with an energy efficiency of $\eta = 0.4$, corresponding to a reversibility of $\epsilon = 0.65$ (indicated by the lower cross (HW) in Figure 6).

The considerations above indicate the importance of carrying out the gasification process as far as possible without losses. For allothermal processes, in which external heat is added, not only is it possible for heat losses to be reimbursed but also more water can be fed into the gasification process. Thus, the hydrogen content of the product gas can be raised considerably (see the upper line in Figure 5). The additional heat can be used directly and completely to "split" water. This heat has to be provided at the temperature of the molten-iron bath, 1,600 K, and has to be introduced into the bath at a considerable power density, 3 to 5 MW/m^2 of molten-iron bath. In the near future it does not seem likely that nuclear or solar heat will fulfill this requirement directly. Nevertheless, such a possibility is indicated by the dotted line in Figure 7 and the resulting large saving in energy (lossless allothermal gasification) is visible in Figure 8, where the dotted line represents the total energy expense. The upper section beneath this line shows the fraction of 1,600 K heat, the middle section indicates the amount of 800 K heat (for electrolysis), and the base section gives the coal requirement. For the example coal of composition CH ($\nu = 1$), it is necessary to add only 0.77 units of heat (0.54 units at 800 K and 0.23 units at 1,600 K) to the 0.75 units of coal of composition CH, where 0.75 represents the chemical minimum. The overall energy efficiency, therefore, is $\eta = 0.66$.

This saving in energy, resulting from the energetically "cheap" production of hydrogen through the admission of heat of temperature 1,600 K into the

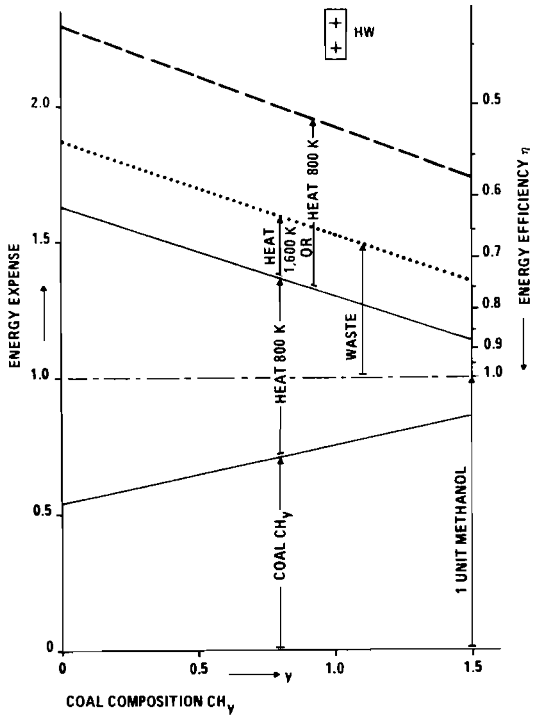


FIGURE 8 Energy expense (800 and 1,600 K heat and coal) required to produce 1 unit of methanol, as a function of coal composition CH_y, for lossless allothermal gasification. HW: Humboldt-Wedag gasification, $\cdots\cdots$ 800 K heat + 1,600 K heat + coal, - - - - 800 K heat + coal.

gasification process is desirable. Since there seems to be no possibility, at present, that this heat could be provided directly by nuclear or solar energy, it might be suitable to introduce a type of "heat pump" into the process. Certainly no heat pumps in the conventional sense exist for such high temperatures. However, the combination of a thermal power plant (supplied with heat of temperature T_1 and dissipating heat at T_0) and electrical heating at temperature T_2, with $T_1 < T_2$, is indeed a form of heat pump, although not a reversible one. At present, for T_1 = 800 K, the efficiency of electricity generation is 0.4, and the efficiency of electrical heating can be taken as $\eta = 0.9$ at 1,600 K (inductive, arc, or resistive heating), therefore the overall energy efficiency of such a heat pump is $\eta = 0.36$. This corresponds to a reversibility of $\epsilon = 0.47$, which is a reasonable figure and comparable to that of conventional heat pumps and cooling equipment. The figure $\epsilon = 0.47$ is based on the energy efficiency of a reversible process, which is supplied with heat of temperature T_1 = 800 K, yields heat of temperature T_2 = 1,600 K, and dissipates heat at T_0 = 300 K, thus $\eta_{rev} = [(800 - 300)/800] [1{,}600/(1{,}600 - 300)] = 0.77$. Such a means

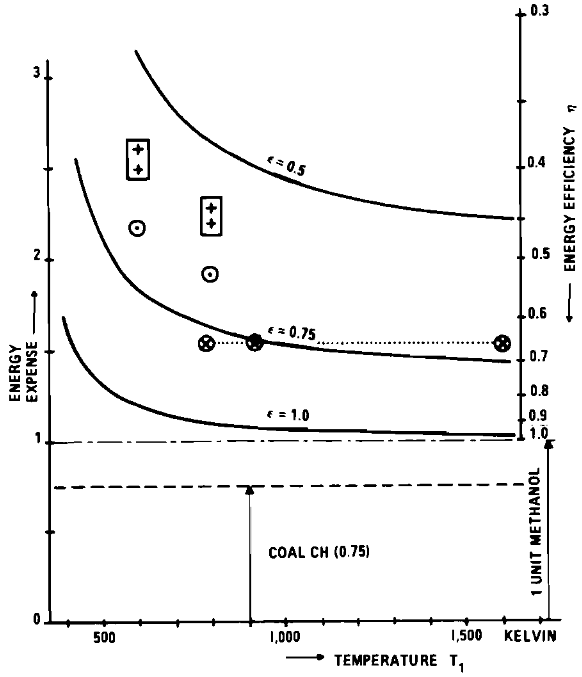


FIGURE 9 Energy expense for 1 unit of methanol for reversible ($\epsilon = 1$) and irreversible ($\epsilon < 1$) processes. ⊞ , Humboldt-Wedag + power plant ($\epsilon = 0.64$) + electrolysis ($\epsilon = 0.75$); ⊙, Lossless autothermal + power plant ($\epsilon = 0.64$) + electrolysis ($\epsilon = 0.75$) or lossless allothermal (electr. heated, $\eta = 0.9$) + power plant ($\epsilon = 0.64$) + electrolysis ($\epsilon = 0.75$); ⊗ , Lossless allothermal (high temp. heat) + power plant ($\epsilon = 0.64$) + electrolysis ($\epsilon = 0.75$).

of providing 1,600 K secondary heat from 800 K primary heat is indicated by the broken line in Figure 7, and the total energy expense is represented by the broken line in Figure 8 (for varying compositions of coal). This energy expense, however, amounts to the same as for the autothermal process, hydrogen being electrolytically produced to compensate for the hydrogen lacking in the synthesis gas (Figure 6). Thus, the overall result for the allothermal and autothermal processes is the same. This result must not be regarded as negative. It indicates that the choice between the two processes is not restricted by energetic considerations since in this respect the processes are comparable, but it can instead be based on technical and economic factors.

The reversibility for these examples is shown in Figure 9, where again the energy expense is plotted against the temperature T_1 of the heat expended in the case of coal of composition CH ($y = 1$). Curves of constant reversibility ϵ are given. Detailed energy and exergy flows, energy efficiency η, and reversibility ϵ for the main subsystems are put together in Figure 10 for the configuration proposed as the reference case. For gasification and electrolysis, the

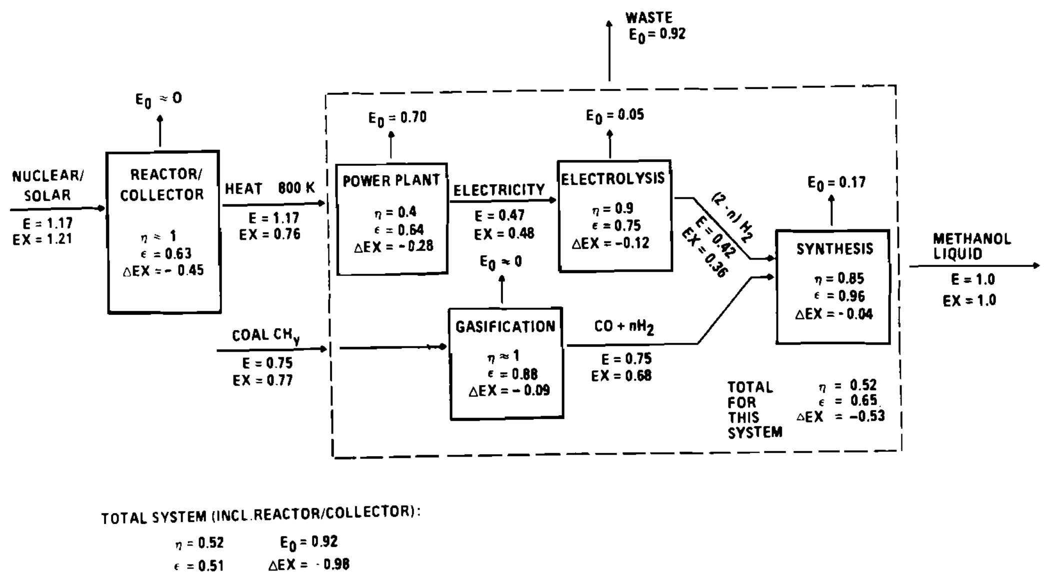


FIGURE 10 Energy (E) and exergy (EX) flows in a methanol production plant (reference case).

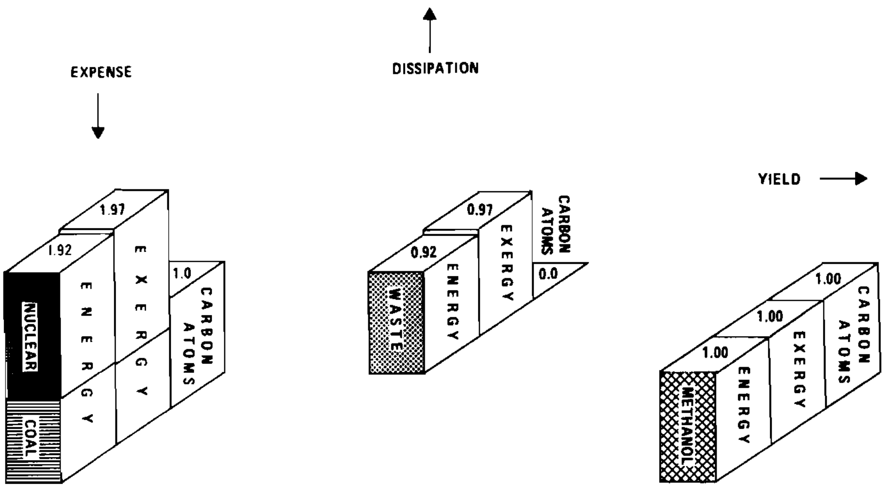


FIGURE 11 The expense of energy, exergy, and carbon atoms required for the production of 1 unit (respectively) of methanol from coal and nuclear energy (reference case).

figures are optimistic but not unrealistic; however, those for the power plant and the synthesis process are conservative.

For the sake of completeness, the production of the expended heat is taken into consideration (Figure 10, left-hand side). For simplicity, the energy efficiency is taken to be $\eta = 1$ for producing the heat by means of a nuclear reactor or a solar collector (for a nuclear reactor and for a concentrating mirror system, this assumption is almost valid). In the reference case, $T_1 = 800$ K, the reversibility for this heat production from primary high quality energy is $\epsilon = 0.63$.

With regard to the expense of primary energy – coal of composition CH and nuclear or solar energy – the total methanol production plant has overall energy and exergy efficiencies of about 0.5. By itself this result is not exciting, but when considered in conjunction with the fact that only the minimum of carbon atoms are used, it appears a relatively attractive means of producing a substitute for fossil oil. Over the long term, improvements in electricity generation are to be expected until methanol is produced on a large scale. To speculate (we will not argue about details), either the temperature could be raised considerably (high temperature reactor) or the reversibility of the thermal conversion process could be improved. Here, only the consequences of such an improvement should be mentioned (e.g., a rise in the energy efficiency from $\eta = 0.4$, as in the reference case, to $\eta = 0.5$). The total energy efficiency η would then increase from 0.52 to 0.59, and the reversibility ϵ from 0.51 to 0.58. The entire lossless and reversible electricity generation from nuclear or solar energy would raise both sets of figures to about 0.8.

So, the reference case, as outlined in Figure 10, can be regarded as a real-

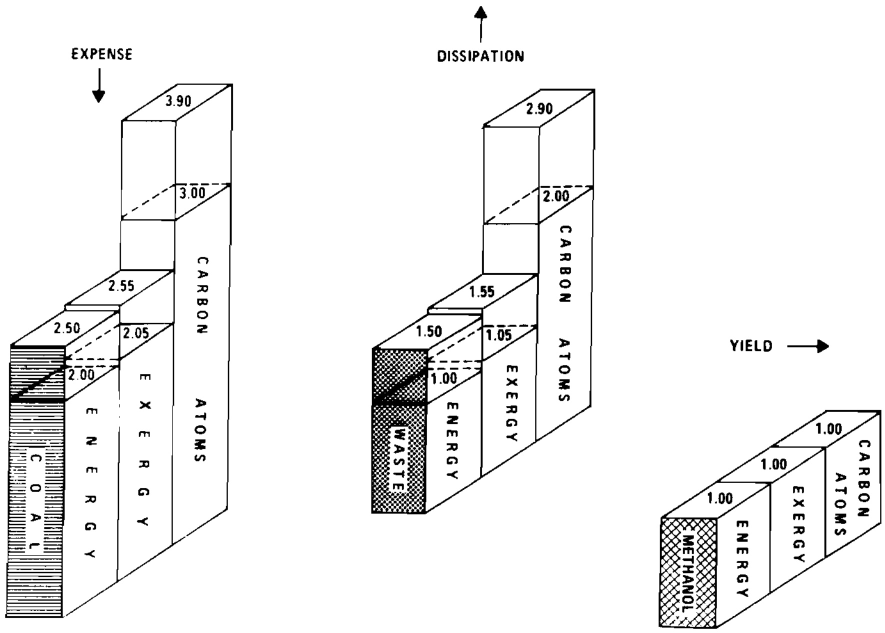


FIGURE 12 The expense of energy, exergy, and carbon atoms required for the production of 1 unit (respectively) of methanol from coal alone (two estimates).

istic technical yardstick by which proposed processes and the development of present processes can be measured. The hypothetical, fully reversible process, $\epsilon = 1$ (the requirements of which have been given in Figure 1), remains the ultima ratio.

The expense, dissipation, and yield of energy and exergy and carbon atoms are represented in a simplified form in Figure 11 for the reference case, normalized for yield = 1 for each of these quantities. It should be emphasized that the reference case includes optimistic assumptions about the gasification and hydrogen-generating subsystems that have not yet been proved to be attainable for large-scale technical equipment.

COMPARISON OF ALTERNATIVES

It is beyond the scope of this investigation to collect all attainable data of processes relevant to methanol production that are under development or consideration and to measure the more technically- and economically-based estimates against the reference case given here. Nevertheless, this should eventually be done. In one of the studies being undertaken at IIASA the technical and economic feasibility of using molten-iron bath coal gasification with additional electrolytic hydrogen for methanol synthesis is being examined and will be reported separately. Our reference case has been chosen with special regard to this system.

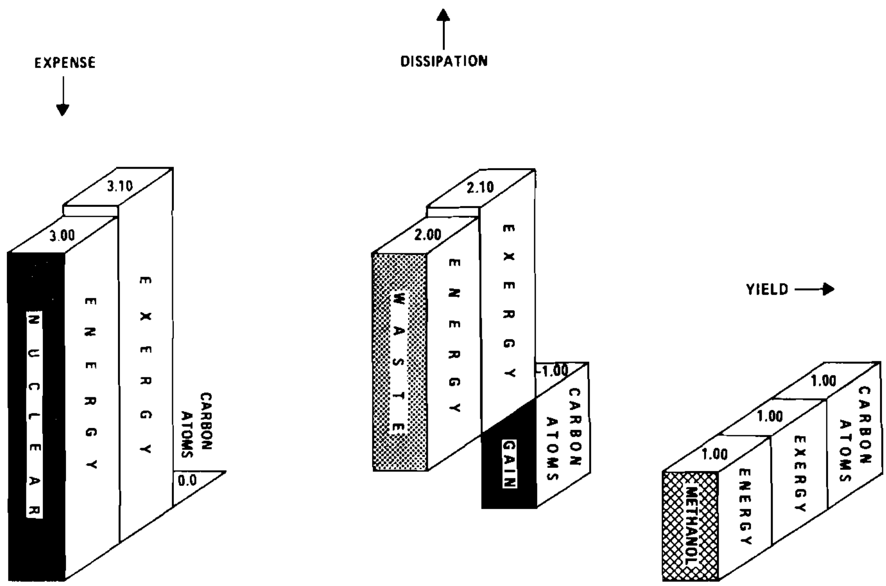


FIGURE 13 The expense of energy, exergy, and carbon atoms required for the production of 1 unit (respectively) of methanol from nuclear energy alone.

At this point two extreme examples of methanol production should be mentioned: methanol produced solely from coal and methanol produced solely from nuclear energy.

In the case where coal is used as the sole source of energy (and of carbon atoms) for methanol production, two governmental studies (Ministry for Research and Technology 1974 and Oversight Hearings 1975) estimate an expense of 2 to 2.5 energy or exergy units of coal for the production of 1 unit of methanol. Waste energy is, therefore, 1 to 1.5 units, and the energy efficiency is 0.5 to 0.4, respectively. However, 2 to 3 carbon atoms have to be dissipated (as carbon dioxide) to gain 1 carbon atom in a methanol molecule (see Figure 12).

In the other extreme case, where nuclear energy is used as the sole energy source, the possibility of extracting carbon dioxide from the air or seawater is considered. Under ideal conditions, the energy expended in separating carbon dioxide from the air amounts to less than 3 percent of the chemical energy of methanol. Therefore, it is not important whether the separation is carried out with a high degree of energy efficiency. The main problem is the considerable size and cost of the facilities required for the separation. Most of the energy expense, however, is necessary for the production of hydrogen since in this case 3 molecules of hydrogen are required for methanol synthesis:

$$CO_2 + 3H_2 \doteq CH_3OH + H_2O \qquad (5)$$

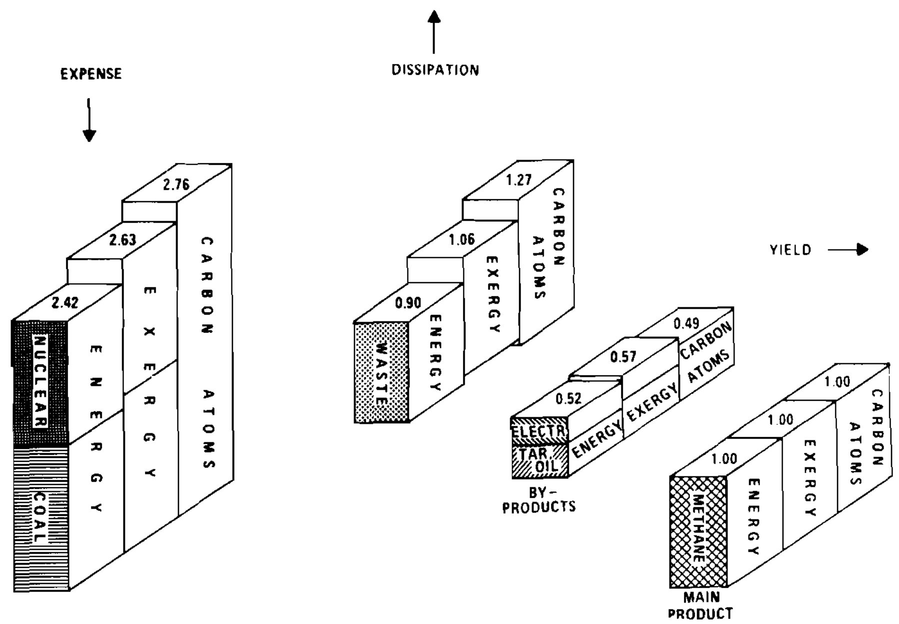


FIGURE 14 The expense of energy, exergy, and carbon atoms required for the production of 1 unit (respectively) of methane from coal and high temperature reactor heat. Source: Nuclear Research Installation (1977).

Estimates of technically feasible processes for producing methanol in this way, therefore, depend strongly on the efficiency of hydrogen production. Overall efficiency rates of between 58 and 94 percent, based on the electrical energy expense, can be expected (Steinberg and Baron 1977). An electrolysis efficiency of 0.9 (a reversibility of 0.75), as in the reference case, would result in an efficiency rate of 83 percent for methanol, based on electrical energy expense. Together with the reference case figure of $\eta = 0.4$ for electricity generation from 800 K heat or from nuclear energy (if this is converted with $\eta = 1$ to heat), we arrive at an overall efficiency of 0.33 for methanol produced from nuclear energy alone. Therefore, from the viewpoint of coal resources and carbon dioxide release, in this attractive option 2 units of heat are dissipated for the production of 1 unit of methanol, but no fossil carbon is used or wasted (see Figure 13). To be more exact, -1 atom of carbon is dissipated (i.e., 1 atom is gained, since it is withdrawn from the air or seawater).

To our knowledge there is only one project at an advanced stage that adds heat from a nuclear source to a gasification process: the "Project Prototypanlage Nukleare Prozesswärme (PNP)", led by an association of German industries and institutions. From 1975 to 1976 basic concepts for different coal gasification processes using heat from nuclear sources (a high temperature reactor) were developed for large-scale plants (Nuclear Research Installation 1977). One of these processes, steam gasification for the production of methane, is represented by its gross balance in Figure 14. For the production of 1 unit of the principal product, methane, about 0.5 to 0.6 units of energy, exergy, or carbon appear in the by-products – electricity, tar, and oil; 0.9 to 1.3 are dissipated in the environment; therefore an expense of 2.4 to 2.8 units of energy, exergy, and carbon is required. Coal and nuclear energy contribute almost equal shares to the energy and exergy expense. This may not appear very satisfactory compared to the reference case, but it has to be taken into account that this project is already at an advanced stage. The detailed planning for a prototype of 750 megawatts (MW) will be completed in 1982, whereas the reference case should be considered as a long-term asymptote.

To return to the initial problem concerning the primary energy requirements for the substitution of methanol for fossil oil over the long term, the answer, in simple and summarized terms, is that:

1 TW methanol requires 0.8 TW coal + 1.2 TW nuclear
or 2 TW coal solely
or 3 TW nuclear solely.

At present, short term requirements are estimated to be 20 to 50 percent larger than those given above.

From this and other aspects such as resources, the environment, economics, market penetration, it is expected that, given our present knowledge, the coal plus nuclear option will be the most attractive, with the possibility of a smooth transition to a solely nuclear option in the future.

REFERENCES

El-Sayed, Y. M., and R. B. Evans (1968) Thermoeconomics and the Design of Heat Systems. Report. Hannover, New Hampshire: Thayer School of Engineering, Dartmouth College.

Getoff, N. (1977) Wasserstoff als Energietraeger. Vienna and New York: Springer.

Haefele, W., and W. Sassin (1977) A future energy scenario. Energy Resources: Availability and Rational Use. Digest of the Tenth World Energy Conference, edited by L. Grainger. Guildford, Surrey: IPC Science and Technology Press.

Ministry for Research and Technology (1974) Neuen Kraftstoffen auf der Spur. Bonn.

Nesterov, I. I., and F. K. Salmanov (1977) Present and future hydrocarbon resources of the earth's crust. Future Supply of Nature-Made Petroleum and Gas: Technical Reports, edited by R. F. Meyer. Oxford: Pergamon Press.

Nuclear Research Installation (1977) Prototypanlage Nukleare Prozesswärme. Statusbericht zum Ende der Konzeptphase, Ergebnisbericht der Planungs-, Forschungs-, und Entwicklungsarbeiten. Julich, FRG.

Oversight Hearings, Methanol Derived from Fossil Fuels (1975) Hearings before the Subcommittee on Energy Research, Development, and Demonstration (Fossil Fuels) of the Committee on Science and Technology, U.S. House of Representatives, Vol. I, No. 43, Washington, D.C.

Steinberg, M., and S. Baron (1977) Synthetic carbonaceous fuel and feedstock using nuclear power, air and water. International Journal of Hydrogen Energy 2:189–207.

Voigt, H. (1978) Evaluation of Energy Processes Through Entropy and Exergy. RM-78-60. Laxenburg, Austria: International Institute for Applied Systems Analysis.

MARKET SUBSTITUTION MODELS AND ECONOMIC PARAMETERS

Bernard I. Spinrad

PREFACE

Market penetration by new technologies is an established fact. The curves of penetration obey simple mathematical rules and fit past experience very well. However, it has not been possible to argue rigorously that future market penetrations would follow the same rules, because a theoretical basis for these rules was lacking.

In a 1977 IIASA report (RR-77-22), Peterka proposed such a basis for centrally planned economies; it followed from detailed consideration of their investment practices. Thus, there remained a need for a model that would be heuristically reasonable for market economies. This report explores two such models.

The work reported here provides a basis for including market penetration considerations in the research activities of the IIASA Energy Systems Program. In particular, it has been used in constructing two reference scenarios for 1975–2030, called "High" and "Low," which are important ingredients in the global energy analysis described in detail in the forthcoming book *Energy in a Finite World.*

SUMMARY

Peterka (1977) has proposed a theoretical economic framework from which the logistic model for market penetration may be derived. His basic equation is consistent with the use of capital charge rates equal to amortization rate plus industry growth rate to determine total costs of a technology and with the use of a price that exactly recovers these costs on an industrywide basis. This formalism is consistent with the practice of centrally planned economies, which use the charge and price rules just set forth.

In addition, the Peterka model can also be interpreted as a strategic principle. Using the principle that the attractiveness of investment is proportional to the degree to which a technology is in use, and also to a figure of economic merit, this paper explores companion models for market economies. The most attractive one, which is called the price model, is derived from a strategic principle that rates the economic attractiveness of a technology in proportion to the inverse of the price that would have to be charged for its product. This judgment of attractiveness of the model is based on synthetic problems simulating market substitution in the electric utility industry.

The Peterka model and the market model can be expressed in identical mathematical form, so their qualitative features must be similar.

The mathematical form of the combined model is

$$\dot{f}_i/f_i = \gamma_i \sum_j W_j(d_j - d_i)$$

$$W_i = (f_i\gamma_i)/\sum_j f_j\gamma_j$$

where f_i is the market share of a particular technology, d_i is the total production cost, including capital charges and amortization, and γ_i is a constant of the particular technology. In the Peterka model,

$$\gamma_i = 1/\alpha_i$$

where α_i is the specific capital investment per unit of production capacity of technology i. For the price model,

$$\gamma_i = \rho/d_i$$

where ρ is the logarithmic expansion rate of the industry.

Both models are pseudo steady-state models, but all the parameters may be expressed as functions of time without violating the principles of the heuristics on which they are based.

1 THE PETERKA MODEL

The fact of market substitution is well established, as is the generally logistic curve shape of the process [see, for example, Fisher and Pry (1970), Marchetti and Nakicenovic (1979), Nakicenovic (1979), or Fleck (in preparation)]. However, the theoretical basis of logistic substitution has only been suggested. One attempt is that of Peterka (1977), who provides a model in which investment in a technology is made at a rate such that new facilities are financed by the marginal income from existing facilities of the same type. Mathematically, this is expressed as

$$\alpha_i \dot{P}_i = P_i(p - c_i) \tag{1}$$

where P_i is capacity of plants exhibiting technology i, α_i is investment required for unit increase of that capacity, p is price of the commodity, and c_i is operating cost per unit commodity. For example, in electrical generation, P_i might be kilowatts; α_i, dollars/kW, and p and c_i dollars/kW-yr; with $\dot{P}_i$ then being yearly capacity addition rate in kW/yr. The operating cost is defined, according to Peterka, so as to include fixed charges against capital such as those for amortization and taxes, but not charges for profit or for accumulation of new capital by the enterprise. These latter items are, rather, taken up in the term $p - c_i$.

The Peterka model has qualitative features that lead to the logistic curves that are observed for market substitution. It is, therefore, an appropriate model to examine for validation or generalization.

1.1 A More Detailed Statement of the Peterka Model

Better insight into the Peterka model can be gained by a more detailed statement of its fundamental principle. This is that the rate of investment in new construction of facilities of a given type is governed by the cash flow generated by existing facilities. The rate of investment in new construction is not entirely due to expansion but also arises from the need to replace existing plant as it is retired. Thus, this rate of investment is $\alpha_i(P_i + a_iP_i)$, where a_i is the amortization rate. The amortization rate is not, of course, necessarily a constant. It could be very small for a technology that is just beginning to penetrate and that therefore consists primarily of new plants. This point can be of importance, as we shall discuss later. However, for the time being we assume that a_i is constant in time. An approximate justification for this can be made by considering that a_iP_i is an allowance for amortization that is applied to replacement construction as required; then, we interpret the term as a required addition to a sinking fund, which, however, neither pays interest when it goes negative nor receives interest when it is positive, and which averages over the long term to zero value.

The cash flow generated by existing facilities is the difference between income pP_i and costs. These costs consist of:

- Operating costs for labor, materials, fuels, services, and other items purchased in proportion to production rate. The unit operating costs are defined as b_i, and the operating costs are therefore b_iP_i.
- Value-added taxes, which can be expressed as a fraction β of operating costs, or βb_iP_i.
- Regularly assessed capital charges, such as those for dividends and interest in market economies, property taxes, insurance, and maintenance. We lump these under a fixed-capital-charge rate δ, and the costs are $\delta\alpha_iP_i$. Note here that we have assumed that the rate δ is invariant among competitive technologies. This is generally the case as a first approximation, but a detailed treatment would show some variation among δ_i defined for different technologies.

With these qualifications, we may write Eq. (1) in more detailed form as

$$\alpha_i(\dot{P}_i + a_iP_i) = P_i[p - (1 + \beta)b_i - \delta\alpha_i] \tag{2}$$

Transposing the term $\alpha_ia_iP_i$ gives

$$\alpha_i\dot{P}_i = P_i[p - (1 + \beta)b_i - (\delta + a_i)\alpha_i] \tag{3}$$

Equation (3) is identical with Eq. (1) provided that we define

$$c_i \equiv (1 + \beta)b_i + (\delta + a_i)\alpha_i \tag{4}$$

Indeed, Eq. (4) provides a more precise interpretation of "cost."

1.2 Mathematical Inferences from the Peterka Model

If we divide both sides of Eq. (1) by α_i and then sum over i, we can derive

$$\dot{P} = \sum_i [P_i(p - c_i)/\alpha_i] \tag{5}$$

Defining logarithmic expansion rate ρ as

$$\rho \equiv \dot{P}/P \tag{6}$$

and market share f_i of technology i as

$$f_i \equiv P_i/P \tag{7}$$

where P is total production capacity of the industry, we can then get

$$\rho = \sum_i [f_i(p - c_i)/\alpha_i] \tag{8}$$

or solve for the price p as

$$p = [\rho + \sum_i (f_i c_i/\alpha_i)] / \sum_j (f_j/\alpha_j) \tag{9}$$

Thus, the price is fixed within the model by the market shares.

Equation (1) can be expressed in market shares by

$$\begin{aligned} \alpha_i \dot{f}_i &= f_i(p - c_i - \alpha_i \rho) \\ &= f_i \left[\frac{\rho}{\Sigma_j (f_j/\alpha_j)} + \frac{\Sigma_j (f_j c_j/\alpha_j)}{\Sigma_j (f_j/\alpha_j)} - c_i - \alpha_i \rho \right] \\ &= \frac{f_i \, \Sigma_j (f_j/\alpha_j)(c_j + \rho\alpha_j - c_i - \rho\alpha_i)}{\Sigma_j (f_j/\alpha_j)} \end{aligned} \tag{10}$$

where, in deriving Eq. (10), we have used the fact that the sum of the f_i is unity.

The term $(c_j + \rho\alpha_j)$ has the character of an augmented cost, the true cost plus a "profit" required to maintain system expansion. We define this as

$$d_j \equiv c_j + \rho\alpha_j \tag{11}$$

The price can be expressed in terms of d_j as

$$p = \frac{\Sigma_j (f_j d_j/\alpha_j)}{\Sigma_j (f_j/\alpha_j)} \tag{12}$$

and market shares change as

$$\dot{f}_i = \frac{(f_i/\alpha_i)\, \Sigma_j (f_j/\alpha_j)(d_j - d_i)}{\Sigma_k (f_k/\alpha_k)} \tag{13a}$$

$$= (f_i/\alpha_i)(p - d_i) \tag{13b}$$

1.3 Economic Implications of the Peterka Model

The statement of the principle of self-financing of an industry's expansion is implicit in Eq. (1), as a consequence of the detailed balancing of each component technology. As is seen in Eq. (9), there is also an implicit price-setting in the Peterka model, which results in there being no excess "profit" beyond what is needed to finance expansion. Thus, no external funds flow into the industry, nor do funds leave the industry for application to other sectors.

This set of conditions describes in an idealized way the principles of price determination in centrally planned economies, often referred to as Libermanism. The industrial expansion rate replaces the investment charge rate of market economies, and plays the same role as a cost factor. Because flows of capital to and from other parts of the economy are not considered, there is no room for external or for distributed profit. [Peterka does, in fact, exhibit a formalism where extra investment, as is necessary to introduce a technology in the first place, is explicitly included. However, this formalism is not developed; the Peterka model is usually stated as Eq. (1).]

Equation (1) suggests that there is a figure of economic merit by which technologies may be ranked. Those technologies for which $(p - c_i)/\alpha_i$ are greatest grow fastest. This makes intuitive sense. It states that one emphasizes those technologies for which the ratio of cash accumulation to investment is greatest. The principle is plausible for both centrally planned and market economies. However, market economies have a price-setting mechanism different from Eq. (9); further, as we shall see later, the economic assumptions of classical market theory suggest that different figures of merit should be used.

1.4 The Peterka Model as a Strategic Principle

The concept of the ratio $(p - c_i)/\alpha_i$ as a figure of merit invites extension. There must be some one of the technologies for which this figure of merit is a maximum. Why do we not concentrate all new construction on this "best" technology? Fleck (in preparation) has analyzed the decision process as one that involves psychological components and that is essentially stochastic. Simplifying these arguments, one can say that the probability of adopting a particular choice has two components. One of these is a figure of merit, and this has just been noted for the Peterka model. The other is a measure of confidence in the specific choice. There are always "opportunity-conscious" and "risk-averse" decision makers. The most opportunity-conscious decision maker will always choose the option with the highest figure of merit. The most risk-averse decision maker will, on the other hand, always choose the option that is most common at the time of decision – the tried and true, so to speak.

One could also justify the factor P_i on the right-hand side of Eq. (1) in a related, but slightly less psychologically oriented, way. At the time of decision, there is always some uncertainty about achieving the economic performance predicted as the figure of merit. The more experience that exists, the less the uncertainty will be. The reciprocal of uncertainty then measures the confidence that one has in the figure of merit, and this positive attribute increases with P_i.

Thus we can say that the expansion rate of technology i, $\dot{P}_i$, can be considered to be a function of two parameters: economic attractiveness, described by a figure of merit E_i, and confidence in the technology, described by a figure of merit C_i. Most generally,

$$\dot{P}_i = \dot{P}_i(E_i, C_i) \tag{14}$$

where the functional dependence is such that $\dot{P}_i$ increases with E_i, C_i within the domain of realizable systems.

Equation (14) can be explored through examination of a variety of functional relations: additive laws, multiplicative laws, and additions and multiplications of powers of E_i and C_i. Fleck's analysis offers justification for the mathematically tractable simple multiplication law

$$\dot{P}_i = kE_i \cdot C_i \tag{15}$$

and Peterka's model is an expression of Eq. (15) for which E_i is identified with $(1/\alpha_i)[\Sigma_j (f_j d_j/\alpha_j)/\Sigma_k (f_k/\alpha_k) - c_i]$, C_i is identified with P_i, and k solves to be unity.

Equation (15) is the *strategic principle* adopted throughout this paper, and the identification of C_i with P_i is likewise robust. We shall later be examining other figures of merit, believed to be more descriptive of market economies.

Considering Peterka's model to be a strategic principle removes one heuristic objection to it. As pointed out, whatever the price-determination mechanism is, the figure of merit $(p - c_i)/\alpha_i$ is a plausible one. Maximizing the ratio of earnings to investment is in the investor's interest, be the investor a public body or a private one. Logically, this leads to the model set:

$$\begin{aligned} P_i &= \dot{P} \qquad (i = k) \\ &= 0 \qquad (i \neq k) \end{aligned} \tag{16}$$

where k is that technology for which $(p - c_k)/\alpha_k$ is a maximum. The model of Eq. (16) is optimal, but there is considerable evidence that it is incorrect; there are many instances of favorable technologies that were never deployed extensively because they remained "unfashionable" up to the time that the industry to which they pertained declined. Equation (16) also predicts that small changes in d_i over time cause sudden activities of technology, whereas social systems do not easily accommodate to such "bang–bang" control.

2 MARKET ECONOMY MODELS

We have noted that Peterka's model implicitly incorporates a price-setting mechanism that corresponds to a standard practice of centrally planned economies. This arises from the absence of capital flows into or out of the particular industry. In market economies, such capital flows exist and are (ideally) controlled by the market for capital. Thus, we first look for models that differ from Peterka's only by permitting such capital flows.

2.1 Fixed-Price Models

The most direct extension of the Peterka strategy is to retain the figure of merit but to let the price be fixed arbitrarily. That is, as in the Peterka model,

$$\dot{P}_i = kP_i[(p - c_i)/\alpha_i] \tag{17}$$

It is important to note that the term c_i includes, for market economies, interest and fixed dividends to investors, as well as capital taxes. The inclusion of the constant k, from Eq. (16), as an arbitrary normalizer, permits the price, p, to be an extrinsic parameter. We can solve for k by summing both sides of (17) over all i and noting that $\Sigma_i \dot{P}_i = \rho P$. The result is

$$k = \frac{\rho}{\Sigma_i\ [f_i(p - c_i)/\alpha_i]} \tag{18}$$

and leads to

$$\dot{P}_i = \frac{\rho P_i(p - c_i)/\alpha_i}{\Sigma_i\ [f_i(p - c_i)/\alpha_i]} \tag{19a}$$

or, after some manipulation,

$$\dot{f}_i = \rho f_i \frac{\Sigma_j f_j[(p - c_i)/\alpha_i - (p - c_j)/\alpha_j]}{\Sigma_j\ [f_j(p - c_j)/\alpha_j]} \tag{19b}$$

The similarity to the Peterka model is emphasized if, given an arbitrary price p, we define a parameter

$$\lambda \equiv (1/\rho) \sum_j\ [f_j(p - c_j)/\alpha_j] \tag{20}$$

Then, price is expressible as

$$p = \frac{\Sigma_j\ (f_j/\alpha_j)(c_j + \lambda\rho\alpha_j)}{\Sigma_j\ (f_j/\alpha_j)} \tag{21}$$

Instead of "excess profit" $\rho\alpha_j$, this excess profit is $\lambda\rho\alpha_j$. If we consider total charges d_i as incorporating excess profit, we can for arbitrary prices define

$$d_i \equiv c_i + \lambda\rho\alpha_i \tag{22}$$

Algebraic manipulation then leads to

$$\dot{f}_i = \frac{(f_i/\lambda\alpha_i)\, \Sigma_j\, (f_j/\alpha_j)(d_j - d_i)}{\Sigma_k\, (f_k/\alpha_k)} \tag{23}$$

which differs from (13a) only in having the extra divisor λ on the right-hand side. The system behaves exactly as if each specific investment α_i had been arbitrarily renormalized by the factor λ. Notice, however, that the analogue of Eq. (1) is

$$\alpha_i \dot{P}_i = P_i(p - c_i)/\lambda \tag{24}$$

so that if λ is different from unity, we can tell whether the actual cash flow is into the industry ($\lambda > 1$), or out of it ($\lambda < 1$). This situation permits us to define λ as an investment flow parameter.

2.2 *Investment Opportunity Models*

The standard description of investment planning in market economies, and particularly in industrial sectors, is *not* one in which cash flow per unit investment is to be maximized. Instead, it is assumed that there exists a "fee" for the use of money, and that any amount of capital is available if that fee is paid. Such "fees" are included in the c_i of market economies. For an investment in a new industry, the fee is the going interest rate, augmented by a (market-determined) rate to accommodate the factor of risk.* The objective is then to maximize earnings over and above that fee (see Riggs 1968 and Massé 1962).

The topic of market penetration assumes an existing industry, so only this case will be treated further.

The existing rate of return, to be denoted by r, is simply the cash flow rate divided by the total investment. Then,

$$r = \sum_j \,[(p - c_j)f_j] \,/ \sum_j d_j f_j \tag{25}$$

We may derive the price from Eq. (25) as

$$p = \sum_j f_j(c_j + r\alpha_j) \tag{26}$$

*This statement is a condition that the *enterprise* does not have the opportunity to invest in other, more profitable industries. One might say that this is the decision of the owners (shareholders). If the opportunity exists for them, capital will flow out of one and into the other until (risk-adjusted) rates of return are balanced – at least under ideal conditions. In a dynamic economy, of course, this balance is hardly ever achieved.

Now suppose the industrial expansion target is taken as some fixed δP. If that δP is constructed using technology i, we will make money at a rate $\delta P(p - c_i)$.

δP is a constant and can be absorbed into the constant k of Eq. (15). "Excess earnings" as a figure of merit then leads to

$$(p - c_i) = E_i \tag{27}$$

Assuming as usual that C_i is given by P_i, Eq. (15) becomes

$$\dot{P}_i = kP_i(p - c_i) \tag{28}$$

Summing both sides and expressing the result in terms of the f_i gives

$$\rho = k(p - \sum_i f_i c_i) = kr \,\Sigma\, f_i \alpha_i \tag{29}$$

This development leads to

$$k = \rho/(r\,\Sigma_i\, f_i \alpha_i) \tag{30}$$

and to

$$\dot{f}_i = \frac{\rho f_i (p - c_i)}{r\,\Sigma_i\, f_i \alpha_i} - \rho f_i \tag{31}$$

which in turn reduces to

$$\dot{f}_i = \frac{\rho f_i\, \Sigma_j\, f_j (c_j - c_i)}{r\,\Sigma_j\, f_j \alpha_j} \tag{32}$$

Equation (32) is a close analogue of the Peterka model. The economic differences are the following. First, the availability of capital, in large amounts at a standard rate, has the effect of averaging specific investment as an inhibiting factor; α_i is replaced by $\bar{\alpha}$. Second, the c_i already includes interest and normal dividends on capital investment, and in this sense is analogous to the Peterka model's d_i. Third, the ratio of expansion rate to excess rate of return is a specific accelerator for market substitution. These differences all seem heuristically plausible and make Eq. (32) a candidate for the desired market-economy analogue of the Peterka model.

2.3 A Price-Suggested Model

Equation (32), while plausible, is not quite satisfactory, because the multiplier $1/\bar{\alpha} = 1/\Sigma_j\, f_j \alpha_j$ is on the right-hand side. If we argue that the availability of capital is not a basic problem in market economies (that only the *cost* of capital must be considered), this term, which has the force of an accelerator of technological change, is not heuristically consistent.

Therefore, we look further for a new model. We define

$$d_i \equiv c_i + r\alpha_i \tag{33}$$

We note the identity

$$\sum_i P_i p = Pp = \sum_i P_i d_i \tag{34}$$

If we differentiate (34) with respect to time, we get

$$\sum_i \dot{P}_i d_i = p\dot{P} + P\dot{p} \tag{35}$$

By using the theory of price–demand coefficients and the definition of ρ, we could express the right-hand side of Eq. (35) as a constant multiplied by pP; but that is unnecessary. The important point to note is that the right-hand side is not a function of i.

Now, we note that the Peterka model derives its basic weighting from the appearance of a term $\Sigma_i\ \alpha_i \dot{P}_i$ in the equation describing investment rate balancing. Applying the same reasoning as that model, we get

$$E_i = 1/d_i \tag{36}$$

when we consider income balancing. The figure of merit is the reciprocal of the cost, computed at the rate of return of capital for the industry. There is no need for any other factor in E_i, since any constants from the right-hand side of Eq. (35) can be absorbed into the k of our strategic model, Eq. (15).

The necessary algebra then gives us

$$k = \rho / \sum_j (f_j/d_j) \tag{37}$$

and

$$\dot{f}_i = \frac{(\rho f_i/d_i)\ \Sigma_j\ (f_j/d_j)(d_j - d_i)}{\Sigma_k\ (f_k/d_k)} \tag{38}$$

Equation (38) cannot be proven to be the best analogue of the Peterka model for a market economy, but it seems to be free of the heuristic objections raised against Eq. (32).

3 COMPARING THE MODELS

We display again the models that are favored, in their market share form:

Peterka, Planned Economy, Eq. (13a)

$$\dot{f}_i = \frac{(f_i/\alpha_i)\ \Sigma_j\ (f_j/\alpha_j)(d_j - d_i)}{\Sigma_k\ (f_k/\alpha_k)}$$

Cost Model, Market Economy, Eq. (32)

$$\dot{f}_i = \frac{\rho f_i \, \Sigma_j \, f_j(c_j - c_i)}{r \, \Sigma_k f_k \alpha_k}$$

Price Model, Market Economy, Eq. (38)

$$\dot{f}_i = \frac{(\rho f_i / d_i) \, \Sigma_j \, (f_j / d_j)(d_j - d_i)}{\Sigma_k \, (f_k / d_k)}$$

They all can be expressed in a common form:

$$\dot{f}_i = \gamma_i f_i \sum_j W_j(e_j - e_i) \tag{39a}$$

where W_i are weighting factors defined by

$$W_i = \gamma_i f_i / \sum_k \gamma_k f_k \tag{39b}$$

The parameters are different for the three cases, however.

In the Peterka model,

$$\gamma_i = 1/\alpha_i \; ; \qquad e_i = d_i = c_i + \rho\alpha_i \tag{40a}$$

In the cost model,

$$\gamma_i = \rho/(r \, \Sigma_j \, f_j \alpha_j) = \rho/(r\bar{\alpha}) \; ; \qquad e_i = c_i \tag{40b}$$

In the price model,

$$\gamma_i = \rho/d_i \; ; \qquad e_i = d_i = c_i + r\alpha_i \tag{40c}$$

The analogy between the e_i in the Peterka model and the price model is notable, and we have already observed that this makes the price model a more desirable market economy analogue of the Peterka model than the cost model would be.

3.1 Comparative Behavior of the Models

We can get considerable insight into the comparative behavior of the models simply by examining the γ_i. This parameter is essentially an acceleration parameter for technological substitution: it is a factor in the weights W_j and a separate factor in the equation for $\dot{f}_i$.

In the Peterka model, $\gamma_i = 1/\alpha_i$, or, as I prefer to write it, $\gamma_i = \rho/(\rho\alpha_i)$. Regardless, it is clear that, in this model, technologies of high capital cost are inhibited.

In the cost model, $\gamma_i = \rho/(r\bar{\alpha})$. There is no longer any specific inhibition of technologies of high capital cost, but, because of the factor $1/\bar{\alpha}$, capital-intensive industries are inhibited in their rates of technological change. In addition, the factor ρ/r suggests that industries expanding faster than their rate of return will exhibit more rapid market replacement than those for which the converse is true.

In the price model, $\gamma_i = \rho/d_i$. We note that on the average $d_i > r\bar{\alpha}$, so that market replacement will be slower in the price model than in the cost model.

3.2 *Amortization Revisited*

Both market economy models predict that a no-growth industry will be technologically stagnant. That is, the market share of competing technologies will not change with time. The Peterka model predicts very rapid penetration of low-operating-cost technologies under these conditions. Heuristically, we expect changes to occur, even in a no-growth industry. For this to be within the scope of models, we must now examine amortization more carefully in the market models. (The treatment presented in the Peterka model requires, however, no elaboration.)

There are actually two separate effects of amortization in a market economy. One is to impose an amortization charge on the existing capital plant, to take into account the (financial) decrease of plant value over its lifetime. The other is to require new construction as old plant is retired. The financial amortization charge $a_i\alpha_i$ is included in the c_i in market economies. However, the rate of new construction is altered from $\dot{P}$ to $\dot{P} + \Sigma_i a_i P_i$ as well. This has the effect of changing our strategic model (15) to

$$\dot{P}_i + a_i P_i = kE_i C_i \tag{41}$$

Without following through the details, the cost model (32) then becomes

$$\dot{f}_i = \frac{f_i}{r\,\Sigma_j\,\alpha_j f_j}\left[(\rho + \sum_k f_k a_k)\sum_j f_j(c_j - c_i) + r\sum_k \alpha_k f_k \sum_j f_j(a_j - a_i)\right] \tag{42}$$

For the standard case, in which all the a_i are the same, this reduces to

$$\dot{f}_i = \frac{(\rho + a) f_i\,\Sigma_j\,f_j(c_j - c_i)}{r\,\Sigma_j\,\alpha_j f_j} \tag{43}$$

where a is the (common) value of all the a_i.

The price model reduces similarly to

$$\dot{f}_i = f_i\left[\frac{\rho + \Sigma_j\,f_j a_j}{d_i\,\Sigma_j\,(f_i/d_i)} - \rho - a_i\right] \tag{44}$$

and, for constant a, to

$$\dot{f}_i = \frac{(\rho + a) f_i\,\Sigma_j\,(f_j/d_j)(d_j - d_i)}{d_i\,\Sigma_j\,(f_j/d_j)} \tag{45}$$

In other words, the existence of amortization has the effect of permitting penetration of a technology into any industry where new construction is justified. Only when the industry is declining at the amortization rate or faster is market substitution entirely inhibited.

3.3 Comparative Calculations

For comparing the three models, a set of calculations was run on a synthetic case suggested by the structure of the investor-owned electrical utility system of the United States. Three competing technologies were examined simultaneously:

1. A (relatively) low-capital-cost, high-operating-cost technology
2. A higher-capital-cost, lower-operating-cost technology
3. A very-high-capital-cost, very-low-operating-cost technology

These may be thought of, qualitatively, as resembling fossil-fueled generation, nuclear generation, and solar power, respectively. Indeed, an estimate of actual costs of these types of generation (Spinrad 1980) was used to derive initial values, but the numbers were altered considerably in order to examine cases that had variable penetration.

The costs are given in Table 1 for the cases considered. They correspond in terms of charge rates to inflation-free conditions in the United States. However, an excise tax has been added, which is not common in the United States. It corresponds to mild encouragement to conserve energy.

All calculations reported here were made on market economy assumptions. That is, the term in δ, capital charge rate, and so on, in Eq. (4) explicitly includes dividends and interest.

Actual costs and prices, given in terms of dollars per kilowatt-year of electricity, are presented in Table 2 for the three technologies under the economic assumptions of Table 1. It can be seen that Technology 2 is the cheapest in terms of cost, but that as extra capital charges are added, it gives way in terms of price to Technology 1. Technology 3 is also cheaper than Technology 1 in terms of cost, but its price escalates even faster than that of Technology 2 as additional capital charges are assessed. These additional capital charges are ρ, the industrial growth rate, in the Peterka model, and r, the excess return on capital, in the other models.

The various formulae were approximated by year-by-year difference equations. No problems were encountered in the forward integration as long as round-off errors were not allowed to initiate mathematical instabilities in the solutions. This was avoided by renormalizing the sum of the market shares to unity after each integration step.

The case $\rho = 0.025$ corresponds to a stagnant industry (since amortization was not explicitly incorporated into the equations except as a financial charge –

TABLE 1 Economic assumptions used to test market penetration models.

Parameters	Annual rate (%)	Technology No. 1	2	3
Property taxes and insurance	2			
Amortization	2.5			
Dividends and interest	3.5			
Excise taxes on sales	20			
Capital cost[a] [$/kW(e)]		925	1,500	2,000
Operating cost [$/kW-yr]		100	35	6

[a]At design capacity factor. That is, capital cost per unit rating is total capital cost per unit nameplate rating, divided by annual average design capacity factor.

TABLE 2 Costs and prices of power under varying parameter values.

Model and parameter value	Cost or price of power [$/kW-yr] from Technology No. 1	2	3
Cost	174	155	166
Price *Peterka model*[a]			
$\rho = 0.025$	236.55	231	259.20
0.05	264.30	276	319.20
0.075	292.05	321	379.20
Price *Other models*[b]			
$r = 0.01$	219.90	204	223.20
0.02	231.00	222	247.20
0.04	253.20	258	295.20

[a]ρ = Growth rate of industry, fraction per year.
[b]r = Expected excess return on capital, fraction per year.

see Section 3.2). For this case, market shares of Technologies 2 and 3 are listed in Table 3 for the various models and excess capital charge rates used. The initial condition is the set of market shares listed for year 0.

In the Peterka model for this case, there is a slow growth of Technology 2 at the expense of Technology 3, which has a higher-priced product than Technology 1 or 2. Technology 1, which commands a slightly higher price than Technology 2, retains an almost static market share. It is being displaced by Technology 2 at a very slow rate at the end of a 100-year period.

The cost model shows penetration of both Technologies 2 and 3 into the market, at faster rates than Technology 2 penetrated in the Peterka model.

TABLE 3 Comparison of market shares[a] for industrial growth rate $\rho = 0.025$, Technologies 2 and 3.[b]

	Peterka model		Cost model[c]						Price model[c]					
			$r = 0.01$		$r = 0.02$		$r = 0.04$		$r = 0.01$		$r = 0.02$		$r = 0.04$	
Year	No. 2	No. 3	No. 2	No. 3	No. 2	No. 3	No. 2	No. 3	No. 2	No. 3	No. 2	No. 3	No. 2	No. 3
0	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
5	0.129	0.120	0.152	0.132	0.138	0.129	0.132	0.127	0.126	0.125	0.126	0.124	0.125	0.121
10	0.133	0.115	0.184	0.139	0.153	0.132	0.138	0.129	0.127	0.124	0.126	0.123	0.125	0.117
15	0.136	0.110	0.219	0.145	0.168	0.136	0.146	0.131	0.128	0.124	0.127	0.122	0.125	0.114
20	0.140	0.105	0.258	0.149	0.184	0.139	0.153	0.132	0.129	0.124	0.128	0.121	0.125	0.110
25	0.144	0.101	0.299	0.152	0.201	0.142	0.160	0.134	0.131	0.123	0.128	0.120	0.125	0.107
30	0.149	0.097	0.343	0.153	0.220	0.144	0.168	0.136	0.132	0.123	0.129	0.119	0.125	0.103
35	0.153	0.093	0.389	0.153	0.239	0.147	0.176	0.137	0.133	0.123	0.130	0.118	0.125	0.100
40	0.157	0.089	0.435	0.151	0.258	0.149	0.185	0.139	0.134	0.122	0.131	0.117	0.125	0.097
45	0.161	0.085	0.481	0.148	0.279	0.150	0.193	0.140	0.135	0.122	0.131	0.116	0.125	0.094
50	0.166	0.081	0.527	0.144	0.300	0.151	0.202	0.142	0.136	0.122	0.132	0.116	0.125	0.091
55	0.170	0.078	0.571	0.139	0.322	0.152	0.211	0.143	0.137	0.121	0.133	0.115	0.125	0.088
60	0.175	0.074	0.613	0.133	0.344	0.153	0.220	0.144	0.139	0.121	0.133	0.114	0.125	0.085
65	0.180	0.071	0.653	0.126	0.367	0.153	0.229	0.145	0.140	0.120	0.134	0.113	0.124	0.082
70	0.185	0.068	0.690	0.119	0.390	0.152	0.239	0.147	0.141	0.120	0.135	0.112	0.124	0.080
75	0.190	0.065	0.724	0.111	0.413	0.152	0.249	0.148	0.142	0.120	0.136	0.111	0.124	0.077
80	0.195	0.062	0.755	0.103	0.436	0.151	0.259	0.148	0.143	0.119	0.136	0.110	0.124	0.075
85	0.200	0.059	0.783	0.096	0.460	0.149	0.269	0.149	0.145	0.119	0.137	0.109	0.124	0.072
90	0.205	0.057	0.809	0.089	0.483	0.148	0.280	0.150	0.146	0.119	0.138	0.108	0.124	0.070
95	0.210	0.054	0.832	0.081	0.506	0.145	0.290	0.151	0.147	0.118	0.138	0.108	0.123	0.068
100	0.216	0.052	0.852	0.075	0.528	0.143	0.300	0.151	0.148	0.118	0.139	0.107	0.123	0.066

[a] The body of the table consists of market shares for Technologies 2 and 3, rounded off to three decimal places.
[b] Technology 1 has the remaining market share so that the sum is equal to 1.
[c] The parameter r is the excess profit on investment above and beyond normal dividends and interest, expressed as a fraction per year.

The rate is particularly fast for small r. The price model, on the other hand, shows very sluggish market penetration.

The case $\rho = 0.075$ corresponds to a vigorously growing industry. The actual annual growth rate is closer to 5% than to 7.5% since we have not counted the replacement construction required by amortization. The market share evolution for this case, according to various formulae, is given in Table 4.

The Peterka model shows a decline in market share for both Technologies 2 and 3, for which the price in that model is higher than for Technology 1. The cost model shows a very rapid penetration of Technology 2 – the lowest-cost technology – even under the relatively high excess profit margin $r = 0.04$. The price model shows a relatively sluggish growth of Technology 2 for small r, a sluggish decline for large r, and a decline for the high-cost technology, number 3, in all cases.

4 DISCUSSION AND CONCLUSIONS

In trying to model a phenomenon as complex as market substitution, there is no way of ensuring that any algorithm is correct. Instead, all that can be done is to try models out and see whether the results are reasonable. Since "reasonableness" is subjective, there are always grounds for dispute. Yet, from the examples just exhibited, it seems that some models should be preferred.

Specifically, the cost model, Eq. (32), does not lead to results that are easy to justify heuristically. It leads to very rapid market substitutions, even when intuitively one would think that they would be slow – for example, when the industry is stagnant. Further, it is unstable as the excess rate of return r approaches zero.

The price model, Eq. (38), is free of these defects. It gives a completely definite answer as $r \to 0$. It favors that technology which commands the lowest price under market conditions, including whatever rate of return is appropriate. For the examples tested, it shows rather sluggish market substitution, however.

The Peterka model is simpler, but, as has been pointed out, it is based on an assumption that cannot be justified for market economies. This assumption is that the excess rate of return r can be equated with the industry expansion rate ρ. The Peterka model thus penalizes technologies of high capital cost very heavily when an industry is expanding rapidly; yet this is the circumstance under which capital can usually be attracted easily, and large investments can be tolerated if they lead to production economies.

For these reasons, the price model, Eq. (38), seems to be the most sensible starting point for a market-economy analogue of the Peterka model, which is a valid interpretation of the economic protocols of centrally planned economies. An interesting point of departure for future research would be to see how the market penetration process might vary between market and centrally planned economies. Neither of these models, however, can be adopted as more than a suggestion to try, until their correlation with reality is well checked. The

TABLE 4 Comparison of market shares[a] for industrial growth rate $\rho = 0.075$, Technologies 2 and 3.[b]

	Peterka model		Cost model[c]						Price model[c]					
			$r = 0.01$		$r = 0.02$		$r = 0.04$		$r = 0.01$		$r = 0.02$		$r = 0.04$	
Year	No. 2	No. 3	No. 2	No. 3	No.2	No. 3	No. 2	No. 3	No. 2	No. 3	No. 2	No. 3	No. 2	No. 3
0	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
5	0.119	0.106	0.216	0.145	0.167	0.136	0.145	0.131	0.128	0.124	0.127	0.122	0.125	0.119
10	0.113	0.090	0.338	0.155	0.218	0.145	0.168	0.136	0.132	0.123	0.129	0.119	0.125	0.114
15	0.108	0.076	0.475	0.152	0.277	0.151	0.193	0.140	0.135	0.122	0.131	0.116	0.125	0.108
20	0.102	0.064	0.608	0.136	0.342	0.154	0.219	0.144	0.139	0.121	0.133	0.114	0.125	0.103
25	0.096	0.054	0.720	0.115	0.410	0.153	0.248	0.148	0.142	0.120	0.136	0.111	0.125	0.099
30	0.091	0.045	0.807	0.091	0.480	0.149	0.278	0.150	0.146	0.119	0.138	0.108	0.125	0.094
35	0.085	0.038	0.870	0.070	0.548	0.142	0.310	0.152	0.150	0.118	0.140	0.106	0.125	0.089
40	0.080	0.032	0.913	0.053	0.612	0.134	0.344	0.153	0.153	0.117	0.142	0.103	0.125	0.085
45	0.076	0.027	0.942	0.039	0.671	0.123	0.378	0.153	0.157	0.115	0.144	0.101	0.124	0.081
50	0.071	0.022	0.961	0.028	0.723	0.112	0.412	0.152	0.161	0.114	0.147	0.098	0.124	0.077
55	0.067	0.019	0.974	0.020	0.769	0.100	0.447	0.150	0.165	0.113	0.149	0.096	0.124	0.073
60	0.063	0.016	0.982	0.015	0.809	0.089	0.482	0.148	0.169	0.112	0.151	0.094	0.124	0.070
65	0.059	0.013	0.988	0.011	0.842	0.079	0.516	0.145	0.174	0.111	0.153	0.091	0.123	0.067
70	0.055	0.011	0.992	0.007	0.870	0.069	0.550	0.141	0.178	0.110	0.156	0.089	0.123	0.063
75	0.052	0.009	0.995	0.005	0.893	0.060	0.582	0.137	0.182	0.109	0.158	0.087	0.123	0.060
80	0.049	0.008	0.996	0.004	0.913	0.052	0.614	0.132	0.187	0.108	0.160	0.085	0.122	0.057
85	0.045	0.006	0.997	0.003	0.928	0.044	0.644	0.127	0.191	0.107	0.163	0.083	0.122	0.054
90	0.043	0.005	0.998	0.002	0.941	0.038	0.672	0.122	0.196	0.106	0.165	0.081	0.121	0.052
95	0.040	0.005	0.999	0.001	0.952	0.033	0.699	0.116	0.201	0.104	0.168	0.079	0.121	0.049
100	0.037	0.004	1.000	0.000	0.960	0.028	0.724	0.111	0.205	0.103	0.170	0.077	0.121	0.047

[a] The body of the table consists of market shares for Technologies 2 and 3, rounded off to three decimal places.
[b] Technology 1 has the remaining market share so that the sum is equal to 1.
[c] The parameter r is the excess profit on investment above and beyond normal dividends and interest, expressed as a fraction per year.

synthetic problems solved in this report are *not* such a check.

Models with extra free parameters could also be tried. One that is suggested by the behavior of the price model, which exhibits market penetrations that are always in the (intuitively) correct direction, but that are slow, would be to multiply the right-hand side of Eq. (38) by a parameter s. To justify such a parameter, however, one would have to invent a new strategic principle.

ACKNOWLEDGMENTS

The author acknowledges with appreciation a review of an earlier draft of this paper by Dr. V. Peterka, whose constructive comments led to significant revision and improvement, and the help of Dr. G. Krömer in getting the numerical examples into and (correctly solved) out of the IIASA computer.

APPENDIX: An Exact Solution for a Special Case

Peterka has demonstrated that certain features of the solutions to his equations are quite insensitive to the values of the α_i used. From this observation, one derives some interest in the case where γ_i are replaced by constant values $\bar{\gamma}$. The situation is of even greater interest for the price model, as it is even more likely that the $1/d_i$ values will be close than it is that $1/\alpha_i$ will be close – at least for situations where substitution is slow.

If we replace γ_i by $\bar{\gamma}$, the model equations become

$$\dot{f}_i/f_i = \bar{\gamma} \sum_j f_j(d_j - d_i) \tag{A1}$$

This set of equations has a solution in closed form:

$$f_i = \frac{c_i}{\Sigma_j\, c_j \exp[-\bar{\gamma}(d_j - d_i)t]} = \frac{c_i \exp(-\bar{\gamma} d_i t)}{\Sigma_j\, c_j \exp(-\bar{\gamma} d_j t)} \tag{A2}$$

Equation (A2) applies for constant $\bar{\gamma}$, d_i, but it is even more generally

$$f_i = \frac{c_i \exp(-\int_0^t \bar{\gamma} d_i \mathrm{d}t')}{\Sigma_j\, c_j \exp(-\int_0^t \bar{\gamma} d_j \mathrm{d}t')} \tag{A3}$$

when $\bar{\gamma}$ and the d_i vary with time. The c_i are determined, of course, by conditions at the reference time $t = 0$.

REFERENCES

Fisher, J.C., and R.H. Pry (1970) A Simple Substitution Model of Technological Change. Report 70-C-215. Schenectady, New York: General Electric Company, Research and Development Center. See also Technological Forecasting and Social Change 3:75–78, 1971.

Fleck, F. (in preparation) Doctoral Dissertation, University of Karlsruhe.

Marchetti, C., and N. Nakicenovic (1979) The Dynamics of Energy Systems and the Logistic Substitution Model. RR-79-13. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Massé, P. (1962) Optimal Investment Decisions. Englewood Cliffs, New Jersey: Prentice-Hall.

Nakicenovic, N. (1979) Software Package for the Logistic Substitution Model. RR-79-12. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Peterka, V. (1977) Macrodynamics of Technological Change: Market Penetration by New Technologies. RR-77-22. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Riggs, J.L. (1968) Economic Decision Models. New York: McGraw-Hill.

Spinrad, B. (1980) Effects of Accounting Rules on Utility Choices of Energy Technologies in the United States. RR-80-27. Laxenburg, Austria: International Institute for Applied Systems Analysis.

EFFECTS OF ACCOUNTING RULES ON UTILITY CHOICES OF ENERGY TECHNOLOGIES IN THE UNITED STATES

Bernard I. Spinrad

PREFACE

Comparison of the costs of power systems is important: additional money spent by the consumer because he does not use the least expensive system is money that cannot be spent elsewhere. Recognizing this, the IIASA Energy Systems Program has developed scenarios that have, to some extent, matched supply to demand using energy technologies in the order of their economic potential: cheapest ones first. The estimates that have been used for nuclear energy indicate that this method is relatively economic, and therefore should be deployed at an early stage.

However, the cost of nuclear power is a controversial subject. Many estimates of the cost of nuclear power in the United States have been published to demonstrate that this technique is not economic. Since the United States is a major energy consumer, it is necessary to examine the issues more closely.

Costing ground rules that are accepted by both the advocates of nuclear power and their opponents in the United States lead to an apparent dominance of capital charges over fuel-cycle costs. The high capital cost of nuclear systems is therefore the chief reason for claims that they are not economic. Yet, over the lifetime of a nuclear reactor, fuel-cycle expenditures will generally be larger than the capital cost. The cause of this discrepancy seems at first to be inflation, since inflation increases capital charge rates. However, standard methods of accounting take this effect into consideration; in particular, inflationary changes in capital charge rates are matched by equivalent increases in properly inflated and levelized fueling and operating costs for all types of systems.

The source of much of the confusion that appears in comparisons of the cost of nuclear and other power systems, particularly in the United States, must therefore lie in the accounting systems that have been adopted. The effects of inconsistent accounting are examined in this paper.

SUMMARY

Monetary inflation does not change the values of commodities relative to each other, only the value of money relative to commodities. Therefore, it would be expected that a comparison of the cost of technological options would not be inflation-dependent. This is borne out by the fact that when systems are compared using three different methods: (a) reducing all costs to their present worth; (b) reducing all costs to constant-value currency and applying inflation-free discount rates; and (c) levelizing future costs at prevalent discount rates; the same relative cost figures are obtained.

These three methods are used to compare five systems that supply electrical power:

- Light-water reactors (LWR)
- Liquid-metal fast-breeder reactors (LMFBR)
- Coal plants, with scrubbers, burning low-sulfur or processed high-sulfur coal (CS)
- Coal plants, with fluidized-bed combustion of high-sulfur coal (CFB)
- Solar power plants with sufficient storage for base-load use (SS)

Light-water reactors and coal plants with scrubbers are the systems presently in operation, and their "typical" costs can be estimated. Nevertheless, the costs quoted should be considered only as illustrations, since both of these types of plant seem to be subject to potential escalation of capital costs, even in constant dollars. Target costs after development were taken as estimates for the other three systems. Using these data, the cost comparison shows that:

- LWR has a decisive cost advantage over coal
- If target costs are met, LMFBR would be the cheapest system
- If target costs are met, SS is almost competitive with the nuclear systems, and is much cheaper than coal

These conclusions are heretical by currently accepted standards. Spokesmen for US utilities, using cost data similar to those taken here for illustrative purposes, are almost unanimous in their view that coal and nuclear power are closely competitive with each other, and that solar energy is a lost cause even if reasonable cost targets are met. In examining the reasons behind this statement, two major points must be considered. The first is that taxes must be included when comparing prices. Most of these taxes are income taxes, which, because of the capital structure of the utility industry, are effectively capital-cost taxes. This severely penalizes solar power, but does not significantly affect the comparison between nuclear power and coal. The second is that it is common to compare systems using fully inflated capital charge rates, but with operating costs levelized over only a fraction of plant life. This "mixed-mode" accounting does not take the later economic value of the plant into consideration. This economic value depends largely on the recurrent costs, being much higher for plants for which the recurrent costs are low, i.e., nuclear, and especially solar, installations. When uncertainty is considered as a planning factor, however, it is precisely the systems with high recurrent costs that have the greatest likelihood of cost escalation, in an absolute sense. The bias introduced by ignoring the future economic value of the plant is therefore in the wrong direction to counteract the factor of uncertainty.

1 INTRODUCTION

Engineering economics is the art of determining the cost of a manufactured product. To the extent that this determination is correct, the art might also claim to be a science. However, the definition of a "correct" cost has many subjective elements. Even when a plant has been bought for a known sum of money, operating costs are available, and resource inputs can be obtained from a fully developed market, a determination of the overall cost is dependent on future expectations. This arises because capital costs are recovered out of future earnings, and future operating and resource costs affect the expected market for, and value of, the product. The art, therefore, may be described as judicious forecasting of future events, while the science is the use of these forecasts to draw conclusions.

In a period of inflation, the standard forecast is that the cost of purchased goods and services will increase at a constant relative rate in current dollars (i.e., in dollars of account at the time of purchase). For example, if an inflation rate of 6% per year is forecast, steel or wood or bread or wages which cost \$1.00 today will cost \$1.06 one year from now and \$1.79 (1.06^{10}) ten years hence. However, future costs are subject to a discount; the value of a dollar used productively today will increase with time. Inflation is a factor in this discounting, and the contribution of inflation to the discount rate exactly cancels the contribution of inflation to future costs. The result of this procedure is to make the *present worth* of future costs insensitive to inflation; they can effectively be calculated in uninflated, constant dollars. This is both a logical and a conceptually satisfying result, since it eliminates the consequences of a forecast containing an extrinsic factor (the value of money) and essentially puts currency on a "goods and services" basis.

Since the present worth of future expenses can be calculated in a robust fashion, it seems at first glance that the cost of a process can be obtained simply by adding this value to the capital expenses which have been accrued up to the time of plant operation. Capital costs plus present worth of future costs must be covered out of future income. Regardless of how this income is to be realized, the process which has the smallest amount to recover is the cheapest; and other things (e.g., external costs) being equal, the cheapest system is the one to adopt.

However, things are not always as simple as they seem, as we shall see. To understand the reasons for this, it is necessary to recapitulate some of the standard practices in engineering economics.

2 REAL INTEREST

Both classical and Keynesian economics predict that the actual interest rates charged, minus the prevailing rate of inflation, will tend toward a constant value. However, a number of authors have pointed out that changes in the distribution of income, especially between wage-earners and entrepreneurs, can affect the value of this constant. Since distributional changes are functions of the social structure, only long-term trends may be expected to produce a real effect. Dramatic events such as wars, great economic depressions, and very severe inflation would probably cause fluctuations in this value, but secular changes of the basic interest rate might have time constants of the order of 20–30 years: one human generation.

No dramatic events occurred in the economy of the United States between the years 1952 and 1972, i.e., roughly the period between the Korean war and the oil crisis. The end of the period, however, includes the Viet Nam war, which seems to have been financed largely by post-1970 inflation. Figure 1 exhibits the excess of utility bond yields over the rate of inflation in the previous year [taken as the rate of increase of the Consumer Price Index (CPI)] over this period, reduced to "real interest" as a percentage. These data are consistent with the value of 2.75% used by many utility economists to estimate real return on bond offerings. Moreover, the data are also consistent with rates used by both the utilities and the federal government in the 1930s, the only period in the last fifty years during which the CPI remained constant.

Thus, in the rest of this report, the "real" (inflation-free) interest rate will be assigned a value of 2.75% per year, and denoted by I_0.

3 CAPITAL RETURN AND FINANCING CHARGES

Utilities are financed both by borrowing capital at interest, as with bonds, and by selling shares to investors, as with common stock. Since investment carries a risk, the returns made to the stockholders are normally expected to be larger than those made to the bondholders. In the same way that I_0, the inflation-free interest rate, is fixed at 2.75% per year, utility economists tend to use a value of 4% for R_0, the annual stockholder return in the absence of inflation. The difference between I_0 and R_0 is relatively small, as utility investment is considered to be low in risk. A utility is buffered to a certain extent against excessive losses

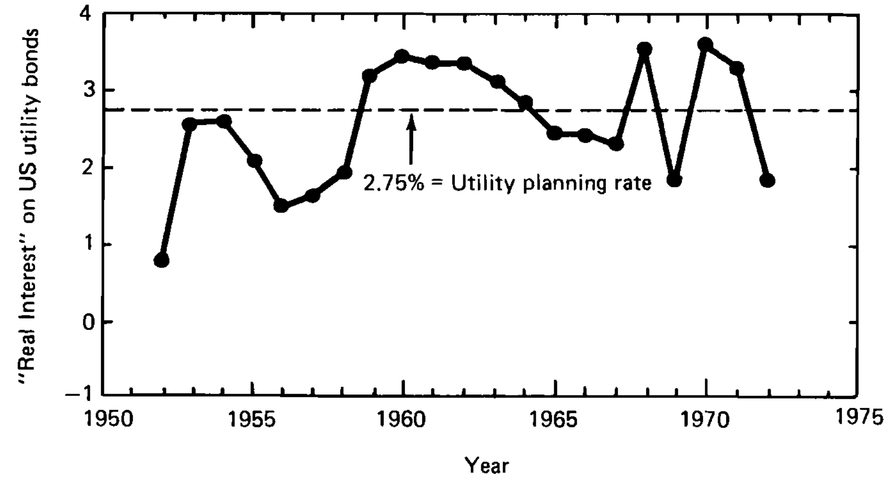


FIGURE 1 Percentage excess of utility bond yields over the rate of inflation in the previous year (as measured by the rate of increase of the Consumer Price Index) for the period 1950–1975 (Statistical Abstract of the United States 1976).

by the monopoly that it enjoys, and prevented from making excessive gains by the compensating regulation of the prices that it can charge.

State regulations, for a variety of reasons, require utilities to raise equity and float bonds at a fixed ratio. While the ratio varies from state to state, it tends to be about 55% stock to 45% bonds. This ratio produces an intermediate value of financial return on capital charges. In the absence of inflation, we can calculate, for $I_0 = 2.75\%$, $R_0 = 4\%$, and an equity: debt ratio of 55:45, the value of the annual financial return, $F_0 = 3.4375\%$.

For planning purposes, then, a utility will consider three factors in estimating costs:

$$\text{Interest rate } I = I_0 + L \tag{1a}$$

$$\text{Investor's return rate } R = R_0 + L \tag{1b}$$

$$\text{Capital finance rate } F = F_0 + L \tag{1c}$$

where L is the expected rate of inflation and I, R, F, L, I_0, R_0, and F_0 are all expressed as a fraction (rather than a percentage) per year. As formulated in Eqs. (1), all rates are taken to be continuously charged.

4 AMORTIZATION

Different plants may have different amortizations, i.e., expected useful periods of service. Amortization cost is computed assuming regular, equal payments to the lender over the period of useful service, i.e., like a mortgage. The yearly payment, under continuous (for example, computed daily) finance charges and payouts is

$$P = DF/[1 - \exp(-FT)] \tag{2}$$

where D is the original capital cost of the plant (e.g., in dollars), F is the finance rate (expressed as a fraction per year), T is the amortization time (years) and P is the payout rate (dollars per year). The present worth of the plant after $t' < T$ years is computed by discounting the payments to be made between t' and T at the finance rate charged. This becomes

$$\begin{aligned} W(t',T) &= \int_{t'}^{T} P \exp(-Ft)\mathrm{d}t \\ &= D[\exp(-Ft') - \exp(-FT)]/[1 - \exp(-FT)] \end{aligned} \tag{3}$$

where $W(t',T)$ is the present worth of an existing plant.

The value of $W(t',T)$ depends on F, the finance rate, which is inflation-dependent. This is not very satisfactory. If F_0 rather than F were used, this external dependence would disappear. Employing Eq. (1c), which relates F, F_0, and the inflation rate, L, and an equation for the present worth of a plant in the absence of inflation, W_0,

$$W_0(t',T) = D[\exp(-F_0 t') - \exp(-F_0 T)]/[1 - \exp(-F_0 T)] \tag{4}$$

the payout rate can be shown to be

$$P' = DF_0 \exp(Lt')/[1 - \exp(-F_0 T)] \tag{5}$$

where P' is the payout rate under these altered conditions. In times of inflation, an " inflation-free" mortgage requires that the yearly payments be the same in terms of constant dollars. The factor $\exp(Lt')$ simply corrects for the shrinking value of the currency of the future.

A mortgage contract that requires payment in equal installments of constant dollars, rather than current dollars, is rare; but this type of arrangement is very useful in dealing with high, and particularly with fluctuating, inflation. A first approximation to such a mortgage is beginning to appear on the home real-estate market: escalating payments are geared to the estimated future income of the mortgagee, a parameter that generally follows inflation quite well. Another, closer approximation to this ideal is the periodic reappraisal by non-regulated industries of their capital assets; capital returns are then based on these reappraised assets (replacement cost accounting).

Under this reasoning, $W_0(t',T)$ as given in Eq. (4) is the correct basis for calculating amortization. Amortization is paid as the difference between the actual regular payments and those that would be made if T were infinite. The rate at which capital is charged for amortization can then be specified, in the absence of inflation, as

$$A_0 = F_0/[\exp(F_0 T) - 1] \tag{6}$$

For $T = 35$ years and $F_0 = 0.034375$ year^{-1}, the amortization rate, A_0, is 0.01475 year^{-1}, which is equivalent to 1.475%.

During inflation, the use of sinking-fund amortization based on the payment of equal installments in current dollars leads to an underestimate of the real rate of depreciation in the (financially) important early years of plant operation. To compensate for this effect, a fictitious amortization time, T', given ideally by

$$T' = \frac{F_0}{F_0 + L} T \tag{7}$$

is sometimes allowed for income tax purposes. The equity of this adjustment has been the subject of much discussion.

5 CAPITAL PAYMENT RATIO

Without any amortization adjustment, the ratio of capital payment with inflation, P', to that without inflation, P_0, is

$$\frac{P'}{P_0} = \frac{F_0 + L}{F_0} \left[\frac{\exp(F_0 T) - 1}{\exp(F_0 T) - \exp(-LT)}\right] \equiv C \tag{8}$$

Table 1 gives the values of the constant C for $F_0 = 3.4375\%$ and various inflation rates, L.

6 LEVELIZED COSTS

In the preceding section it was shown (Eq. 8) that capital charge payments in an inflationary regime are a factor C higher than those in a noninflationary situation. It shall now be shown that the same ratio is also valid for apparent future costs.

In a noninflating economy, recurring expenses costing one unit today will cost one unit tomorrow and so on. In an inflationary economy, these costs increase as $\exp(Lt)$ in current dollars. To evaluate these costs at a constant rate in current dollars, the concept of levelizing is used. In effect, a banking institution acts as the buffer. When the costs are less than the constant amount allocated, the extra money is put into the bank to accumulate interest; when the costs are greater than this allocation, the savings are withdrawn or, if necessary, some money is borrowed. At the end of the period of levelizing, there should be no net credit or debit if the levelized costs have been computed correctly.

TABLE 1 Dependence of the ratio of capital payment with inflation to that without inflation (C) on the annual rate of inflation (L).[a]

L(%)	$C(=P'/P_0)$
0	1.00
2	1.30
4	1.64
6	1.99
8	2.37
10	2.76

[a] F_0 is assumed to be 3.4375% (see Eq. 8).

The present worth of future payments of recurring costs, per unit annual cost, is given by

$$W_1(T) = \int_0^T \exp(Lt)\exp(-F_0 t - Lt)\mathrm{d}t$$
$$= [1 - \exp(-F_0 T)]/F_0 \tag{9}$$

This is the integral of the annual costs, $\exp(Lt)$, multiplied by the discount factor, $\exp(-F_0 t - Lt)$, evaluated over the operating time T. Note that $W_1(T)$ is independent of L.

To levelize future recurring costs, payments must be made at a constant rate in current dollars, such that the present worth is correctly described. We shall call this rate of payment C, for reasons which will become obvious. C can be found from the identity

$$C\int_0^T \exp(-F_0 t - Lt)\mathrm{d}t = W_1(T) \tag{10}$$

Solving for C, the result is Eq. (8). In other words, the ratio of payments made at a constant rate in current dollars to payments made at a constant rate in constant dollars is, for future recurring costs, the same as the ratio of capital payments with and without inflation. Levelizing future costs is therefore a consistent method of current-dollar accounting.

7 PROPER AND IMPROPER ACCOUNTING

Three internally consistent accounting schemes can be used to calculate the cost of making a product. These are:

- *Present Worth.* The properly inflated and discounted costs of future purchases of material and services are combined with the initial capital cost to give the present worth of the entire operation. This method has the advantage that costing can be carried out in either constant or current dollars.

- *Constant Dollar.* This involves simply reducing all payments over time to constant-value currency, and then computing costs according to deflated discount rates. The method has the advantage of requiring no adjustments, but the disadvantage that constant dollars are more often confused with current dollars than vice versa.
- *Levelized Cost.* This is an internally consistent method of accounting in current dollars. It has the advantage that one always knows how many dollars-of-today are to be paid or set aside; the disadvantage is that it creates a somewhat false impression of cost as a function of time.

All three methods will give the same answer to the question "How does the cost of one system compare with that of another?" That is to say, the ratio of the costs of various systems will be the same whichever accounting method is used. The author happens to prefer constant-dollar accounting, for the simple reason that the effects of inflation must be considered at the beginning of the calculation, but this is only a matter of taste.

Constant-dollar and levelized-cost accounting must *not* be mixed, however, since these methods do not express costs in the same units. In fact, levelized costs are not really current-dollar costs, but current-dollar-equivalent costs; the time factor affecting cost is obscured by the levelizing technique. Nevertheless, in inflationary times, one can say that levelized costs are related to constant-dollar costs by the factor C of Eq. (8). In particular, if the capital costs are calculated in current dollars (i.e., with the effects of inflation included in the discount rate), but future costs are put on a constant-dollar (i.e., first-year-cost) basis, the contribution of future costs to the economics of the system is being grossly underestimated. However, this specific misrepresentation is so common in comparing energy systems that it can almost be described as orthodox.

8 ILLUSTRATIVE EXAMPLES

The cost figures for a variety of types of electrical power plant provide a frame of reference for further discussion. It is hoped that the numbers quoted are "realistic," though only in the sense of being typical of expected costs at the time of commissioning. The following systems will be examined:

- Light-water nuclear reactor (LWR)
- Liquid-metal fast-breeder nuclear reactor (LMFBR)
- Coal plants, with scrubbers, burning low-sulfur or processed high-sulfur coal (CS)
- Coal plants, with fluidized-bed combustion of high-sulfur coal (CFB)
- Solar power plants with tower-type collector installation and thermal heat storage (SS)

TABLE 2 Capital costs (1978 $ per kW electric) and economic lifetimes (years) of electrical power plants.

Plant type	Cost	Plant lifetime
LWR	815	35
LMFBR	975	35
CS	550	35
CFB	650	35
SS	2,500 (1923)[a]	70 [35]

[a]Cost less residual value after 35 years, that value discounted to present worth at a rate of 3.4375% per year.

Capital costs for each of these five types of plant are set out in Table 2. In each case, the assumed value is the cost which might be reached *after* full development has taken place. Practically, this means that only the costs for the LWR nuclear and the coal-with-scrubbers (CS) plants are based on working experience. (However, the costs of both coal and nuclear power plants may still be escalating in constant dollars, though in both cases there is scope for technical improvements – and hoped-for improved costs – in what are still immature technologies.) In each case, a plant is to be placed in service in 1978, and the capital cost is expressed in 1978 dollars. All plants have a base-load capacity factor of 65%, although their costs are expressed in terms of nameplate rating. Economic plant lifetimes are also listed in Table 2.

The LWR data of Table 2 are based on midrange values from estimates of LWR costs prepared for the CONAES study of the National Research Council (in press). This procedure led to a basic figure of $675/kW; $150/kW was added to this figure for the cost of the critical reactor core, and $10/kW subtracted for the present worth of residual core value at the end of plant life. The LMFBR cost quoted is, in contrast, a target value. A common target for the capital cost of a developed LMFBR is 1.25 times the cost of an LWR. This leads to a basic figure of $845/kW, to which was added $150/kW for the critical reactor core, and from which was subtracted $20/kW for the present worth of residual core value at the end of plant life.

The CONAES LWR midrange cost is again the reference case for the coal plants, the cost being adjusted on the assumption that coal plants with scrubbers cost 0–20% less than LWRs, the core charges being excluded. This type of plant is therefore costed at 80% of the basic price of an LWR, rounded upward to the nearest $50/kW. For the fluidized-bed plant, however, the value is entirely arbitrary. Many estimates of the cost of developed coal fluidized-bed (CFB) plants predict that the capital costs will be lower than those of coal plants with scrubbers (CS). If this is so, then there is no point in considering coal with scrubbers any further, for, as shown below, the recurring costs of the CFB plant would also be lower.

An estimate of the developed cost of a large solar power installation can only be a guess, but costs of the order of $2,500/*peak* kW have been put forward for desert stations which can be adapted to intermediate load service. This number has been used as it stands, on the basis that the cost of providing thermal storage for conversion of peak capacity to base-load capacity will take up any further economies in plant construction. The number in parentheses is the capital cost corrected for the present worth of plant value after 35 years.

All these plants are large, and large plants tend to be kept in serviceable condition for longer than their conventional write-off time. The estimated life of the solar plant, 70 years, is long enough to be compatible with the idea that it would be superseded only when newer designs requiring less maintenance appear on the market.

Estimates of operating and maintenance (O + M) and fueling (F) costs are listed in Table 3. For the nuclear plants, the fuel costs are those of a fuel cycle with reprocessing, so that LWR and LMFBR can be compared on an equal basis. All steps, including waste management, should be covered under fuel costs. In the case of the two coal plants, however, waste management, including disposal of sludge, is covered under the cost of operating and maintenance.

Again, the costs of operation and maintenance are referenced to the CONAES input numbers for LWR. LMFBR is charged at 10% higher than LWR because of increased plant complexity, CS at 25% higher than LWR due to sludge handling, and CFB (with high-sulfur coal) is strongly penalized for its sulfur-handling needs and consequent high sludge rate. The operating and maintenance cost of solar installations is taken to be half that of LWR plants in view of the smaller work force required.

Light-water reactor fuel cycle costs were calculated using the following data: UO_2 at $100/kg (marginal price); fabricated UO_2 fuel at $100/kg; $100/kg-SWU*; UO_2 reprocessed at $200/kg; waste management fee of $125 per kilogram reprocessed; sales credit at $24/g of plutonium. Advance (or deferred) payments were inflated (or discounted) at 6% per year. Inventory charges are covered under capital costs and are not included here. The use of a 6% discount rate implies constant (1978) dollar accounting within the fuel cycle. The costs are supposed to be those of a fully developed industry.

LMFBR fuel cycle costs were calculated using the following data: fabricated fuel at $800/kg; UO_2 reprocessed at $350/kg; waste management fee of $125 per kilogram reprocessed; sales credit at $24/g of plutonium; and 6% per year escalation or discount rate on payments. Inventory charges are covered under capital costs and are not included here.

The fuel costs of the coal plants are taken to be the costs of coal delivered to the utility, taken here to be in the Midwest of the United States. This is a region of median transportation charges. Typical values are: high-sulfur coal, $1.10 per million Btu (about $30/ton); low-sulfur coal, $1.75 per million Btu

*SWU stands for Separative Work Unit.

TABLE 3 Assumed operating and maintenance (O + M) and fueling (F) costs (1978 mills[a]/kWh) of electrical power plants.

Plant type	Costs		
	O + M	F	O + M + F
LWR	2.1	6.0	8.1
LMFBR	2.3	2.2	4.5
CS	2.6	17.5	20.1
CFB	5.0	11	16
SS	1.0	0	1

[a]One mill is a thousandth part of a US dollar.

(about \$50/ton); plant heat rate, 10,000 Btu/kWh. These numbers have been adjusted for inflation from data presented by Corey (1977) in 1976 dollars, and are marginal costs (i.e., expected costs of new contracts). They are consistent with the data presented by SRI International (1977).

To make the case of solar power comparable to the others, it is necessary to subtract the present worth of the solar plant 35 years from now from the initial capital cost.

The second column of Table 4 lists the sum of present worth of future expenses, plus sunk capital costs, for all the plants considered. A correction for residual assets after 35 years has been subtracted from the present worth of the solar plant. Table 4 therefore shows the total present worth of all payments to be made throughout the lifetime of the installation. As previously noted, all plants are operated at 65% capacity. This amounts to 5,700 hours per year, or 5,700 kWh per year (kW of capacity). The "present worth" column in Table 4 is derived by multiplying the last column of Table 3 by 5.7 to obtain the annual cost (in constant-value dollars) of operating, maintaining, and fueling per kilowatt of electrical plant. This is then converted to the present worth of total expenses over 35 years of operation by multiplying by the factor

$$f = [1 - \exp(-35F_0)]/F_0 \tag{11}$$

and adding the corrected capital costs. The appropriate discount factor is $F_0 = 0.034375$ since the recurring costs have been calculated in constant dollars.

The third column of Table 4 makes the same comparison using the constant-dollar method rather than the present-worth method. Capital charges are taken at $F_0 + A_0$ on the capital costs of Table 2, dividing by 5.7 to convert \$/kWh per year to mills/kWh, and adding the recurring costs listed in the last column of Table 3. The resulting figures are power costs in constant-value dollars. Finally, the fourth column of Table 4 uses the levelized-cost method, assumes

TABLE 4 Comparison of the cost of various types of electrical power plant using three different accounting methods (common financing of capital and recurring costs).

Plant type	Present worth of capital and future costs	Power costs	
	Present-worth method (1978 $/kW)	Constant-dollar method (1978 mills/kWh)	Levelized-cost method – 6% inflation (mills/kWh)
LWR	1,977	15.12	30.15
LMFBR	1,497	12.90	25.73
CS	2,882	24.83	49.52
CFB	2,506	21.60	43.08
SS	2,039	17.57	35.04

6% inflation, a 12.7-year levelizing period, and presents levelized power costs. Note that the ratio of the costs given in the fourth column to those in the third column of Table 4 is numerically 1.99, as given by Table 1 and Eq. (8).

A number of surprising things can be deduced from Table 4. First of all, a comparison of Tables 2 and 4 shows that the present worth of future expenditure exceeds the capital cost of the plant for LWR and both types of coal plant. Secondly, if the solar plant, with storage, were to achieve its objective of an original capital cost of $2,500/kW, it would be competitive with coal, although not with nuclear power. Finally, the cost effectiveness of the breeder reactor looks extremely hard to beat *if its cost objectives are met.*

It must be stressed that this example is strictly illustrative: that is, the input data are purely arbitrary. Another possibility can be examined, as follows:

Retain the light-water reactor as a reference system, but assume that depletion of supplies forces the price of uranium concentrates (in constant-value dollars) to $200/kg, all other cycle costs remaining constant. The fueling cost of the LWR will then rise to 9.1 mills/kWh. Under what circumstances would the other plants be competitive? The capital costs of both types of coal plant are the values given in Table 2; it is then necessary to calculate the price of coal (low-sulfur for the scrubber system, high-sulfur for the fluidized-bed method) that would lead to a cost-based parity between LWR and coal plants. The operating, maintenance, and fueling costs of LMFBR and solar plants are retained at the values given in Table 3; the capital cost required for the systems to be competitive with LWR must then be calculated. The results of these calculations are given in Table 5. It should be noted that this hypothetical case uses a reference price for uranium which is more likely to be representative of the 21st than the 20th century. It is only possible to guess at which of these targets is most likely to be met at that time.

9 INCONSISTENT ACCOUNTING

The results of Tables 4 and 5 are heretical by current standards; yet the same basic data can be manipulated to produce electricity costs that are much more familiar. All that is necessary is to apply the capital charge rate (15–20% per year) used today by the utility industry. Assuming, for illustration, that the capital charge rate is 16% per year, and continuing to use 5,700 kWh per year capacity, the cost of power can be calculated according to the assumptions of Tables 2 and 3. The capital cost (second column, Table 2) must be multiplied by 0.02807 and the result added to the sum of operating and maintenance (O + M) and fueling (F) charges in the fourth column of Table 3. The power costs calculated in this way are given in Table 6. The results of a utility presentation (Corey 1977) for the two cases of available technology are included in Table 6 for comparison.

TABLE 5 Conditions under which various types of plant would be competitive with an LWR fueled with uranium costing $200 per kg (common financing of capital and recurring costs).

Plant type	Variable parameter[a]	Break-even value (1978 $)
LMFBR	Capital cost	1,592 per kW
CS	Delivered cost of low-sulfur coal	31 per ton
CFB	Delivered cost of high-sulfur coal	21 per ton
SS	Capital cost	2,598 per kW

[a]The other parameters remain at the values given in Tables 2 and 3.

TABLE 6 Power costs (1978 mills/kWh) calculated assuming a 16% capital charge rate and present-day fueling costs.

	Power costs			
Plant type	Capital	Recurring	Total	Utility estimate[a]
LWR	22.87	8.1	31.0	28
LMFBR	27.36	4.5	31.9	
CS	15.44	20.1	35.5	40
CFB	18.25	16	34.3	
SS	53.98	1	55	

[a]From Corey (1977).

An examination of Table 6 would lead to the following conclusions:

- The LMFBR has to achieve slightly better values than the targets stated to compete with the LWR
- Coal could compete against uranium at current prices if the delivered price of coal were somewhat reduced
- Solar power is hopelessly expensive

These are, in fact, the general impressions that recur in common small talk, both among people associated with utilities and elsewhere. Is there a fallacy here? And if so, where?

If a fallacy exists, it must be connected with the capital charge rate of 16%. There are, in fact, two fallacies present:

- Confusion of cost with price
- Inconsistent treatment of inflation

The question of price and cost, which involves the differential effects of taxation, will be considered in the next section. In this part of the discussion, however, it is necessary to note that high capital charge rates are always associated with high basic finance rates, and that these high values arise from including inflation in the rates. In other words, they imply that I, R, and F are being used rather than I_0, R_0, and F_0. If actual initial-year operating, maintenance, and fueling costs are being employed, constant-dollar evaluation is necessary. Under these circumstances, inflation should be subtracted from the capital charge rate, i.e., work with I_0, R_0, and F_0. This would greatly reduce the capital charges given in Table 6. Conversely, if capital charges including inflation are used (essentially, current-dollar accounting), then recurrent costs must be levelized. In an inflationary situation, this causes a considerable increase in first-year costs. To summarize, the data of Table 6 are the products of an inconsistent evaluation: one which includes inflation in the capital charge structure, but which also assumes that recurring costs will *not* inflate. It would be an unusual recurring cost that did not inflate with the general economy; indeed, lack of inflation would represent continuing improvements in resource availability and technical economy. Since costs are being estimated on the basis of fully developed technologies, there is no reason to expect that recurring costs would be free from inflation; decisions made only on the basis of short-term charges are thus intrinsically unsound. Table 4 therefore provides a correct reflection of the situation, while the image produced by Table 6 is fundamentally distorted.

10 COST AND PRICE

The customer pays for more than simply the cost of doing business. "Profit" has already been incorporated into the financing factors, F and F_0, but utility income is also heavily taxed, and this tax is added to the price. It is therefore a transfer payment, rather than a simple cost. It is conventional to estimate that taxes are equivalent to the basic return on equity capital, i.e., that taxes represent half of the gross income from equities. Although there are taxes on both property and income, the latter in fact constitute the major share of the tax bill.

No profit is made on amortization. Since recurrent costs are often financed exclusively by debt, they also tend not to be taxed. In fact, under these circumstances the present worths of future operating costs given in the second and last columns of Table 4 have been slightly overdiscounted; Table 7 shows these costs corrected under the assumption of debt-only financing.

The customer contributes to the taxes paid by the corporation in that the prices charged include the effects of this taxation. That the decision is a social one is exemplified by the fact that consumer-owned utilities pay little, or no, tax. Although consumer ownership may be preferred to investor ownership on

TABLE 7 Comparison of the present worth of future expenses and levelized power costs (6% inflation) of electrical power plants – recurring charges financed by debt.

Plant type	Present worth (1978 \$/kW)	Levelized power costs (mills/kWh)
LWR	1,853	32.3
LMFBR	1,551	27.9
CS	3,125	51.9
CFB	2,700	45.4
SS	2,051	38.9

purely ideological grounds, it is nevertheless true that taxes are ultimately a payment to society as a whole for the general services provided to the citizenry. These services are also available to corporate bodies, public and private, and it would therefore seem fair that consumer-owned utilities should also pay taxes. This just serves to illustrate the arbitrary nature of the way in which taxes are levied. Moreover, the appropriate returns from taxes raised from utilities are the identifiable social (external) costs of generating the power, plus the unidentified services which should be allocated to the quantity of power generated. In other words, a combination of excise and value taxes would seem to be more appropriate to the electricity-generating industry than the present taxes on distributed corporate earnings (which are further taxed at the level of the investor's income).

The philosophy of taxation could be discussed indefinitely. Recognizing that taxes are not costs, taxation is not initially considered in the internal planning of the utility. In effect, the utility takes the position that it is merely a collection agent, transferring taxes paid by the consumer to the taxing authority. But at a higher level of corporate planning, taxes must be considered; for the *price* of electricity, which includes taxation, is one of the major factors determining system growth. System growth, in turn, is one of the most important aims of the utility, as this growth tends to make the equity associated with the industry more valuable (i.e., increases the value of stock, *ceteris paribus*), diverting profit for investors into less-taxed capital gains and permitting more self-financing of further investment.

The concern of the utility with prices also cancels out quite effectively any incentive to adopt technologies with high capital costs and correspondingly large investor profit (recalling that profit is made only on capital investment). Since the profit per unit investment is regulated, consideration of the effect of taxes on prices leads to a preference for low-capital technologies.

It may therefore be concluded that planning of utilities is based on the price paid by the consumer, and that the technology with the lowest price will be

TABLE 8 Components of the prices charged by utilities, under constant-dollar and levelized-cost accounting.

Component	Constant dollar	Levelized for 6% inflation	Description
Capital cost	$F_0 = 3.4375\%$	$F_0 + L = 9.4375\%$	Discount rate for capital expenses
Interest	$I_0 = 2.75\%$	$I_0 + L = 8.75\%$	Discount rate for recurring expenses
Amortization	$A_0 = 1.475\%$	$A = 0.3602\%$	
Taxes	$0.55R_0 = 2.2\%$	$0.55(R_0 + L) = 5.5\%$	
Capital + Amortization + Taxes	7.1125%	15.2977%	Capital charge rate against price

TABLE 9 The price of power obtained from different types of plant compared using constant-dollar and levelized-cost accounting. (The cost assumptions are those given in Tables 2 and 3.)

	Constant-dollar price (1978 mills/kWh)			Price levelized for 6% inflation (mills/kWh)		
Plant type	Capital	Recurring	Total	Capital	Recurring	Total
LWR	10.17	8.1	18.3	21.87	16.71	38.6
LMFBR	12.17	4.5	16.7	26.16	9.28	35.4
CS	6.86	20.1	27.0	14.76	41.46	56.2
CFB	8.11	16	24.1	17.44	33.01	50.5
SS	24.00	1	25.0	51.61	2.06	53.7

chosen. Table 8 presents the components of the price calculated using the constant-dollar and levelized-cost methods. Table 9 compares the price of power obtained from different plants under the cost assumptions of Tables 2 and 3, using the same self-consistent accounting techniques as in Table 8. A comparison of Tables 7 and 9 shows that the effect of levying taxes on capital alone is most important for the solar plant. In this case the very high capital cost of the plant makes the taxation burden unusually severe.

11 IMPACT OF UNCERTAINTIES

An argument frequently used in defense of short-term planning horizons is that the future is uncertain. Therefore, it is argued, sunk costs should be recovered as quickly as possible, since the net effect of future uncertainty is to increase investor risk, and this risk has a price. This argument contains some truth, and is the strongest point in favor of the adoption of current-dollar, levelized-cost accounting by utilities. Using this method, capital is actually recovered in the first few years, since the present worth of payments to be made in the distant future is very small.

However, uncertainty of inflation also has its price. If inflation stops, the market value of existing utility bonds increases. If inflation accelerates, the old bonds decrease in value. Utilities can cushion the impact of these changes to a certain extent by refinancing (usually with penalties), while large-scale investors can achieve the same effect using tax allowances. Nevertheless, there is still a financial risk, and the higher the inflation rate, the greater is the risk. The situation with regard to equity is similar, with inflation certainly adding risk to the equity (stock) market.

The investor must respond to this increased risk, and there are some signs that he does so. Figure 1 could be interpreted as an indication that the excess of bond rate over inflation rate rises slightly in a period of high inflation; the "real interest" rate clustered around 2% in the low-inflation 1950s and around 3% in the early and late 1960s when the rate of inflation was higher. However, this tendency is not excessively marked. At most, the real interest rate might have increased by 0.1–0.2% per year for each increase of 1% per year in the inflation rate; however this inference is not statistically strong, and the tendencies noted might have had other causes. The most sound conclusion is that the financial risk associated with fluctuating inflation rates has only a slight influence on financing charges, and that the effects are most probably similar to those produced by increasing the discount rate and the capital charge rate by the same amount.

Any change in the discount rate should also affect future operating costs. The usual estimate of the rate of increase of the recurring costs remains unchanged, however, and to this extent uncertainty *does* require some incremental discounting of future expenses.

The preceding analysis considers only the effects of inflation. What about other uncertainties? It is clear that the estimation of future operating or recurrent costs is more uncertain than that of capital charges and costs. In addition to the fundamental uncertainty of inflation, there are likely to be changes in technology, resource availability, demand, and input-values (e.g., the intrinsic value of labor) which will affect future costs. Predicting these changes is a matter of considerable uncertainty, the uncertainty increasing as the range of the forecast increases.

A qualitative feature of this type of uncertainty is that it tends to be asymmetric. There are always more reasons for increasing real costs than for decreasing them – at least, for operating costs associated with large capital investments. The (Bayesian) curve giving the probability of the recurrent cost being correctly predicted, as a function of the predicted cost, becomes more and more skewed as time goes by (see Figure 2): while the mode tends to remain fixed at a

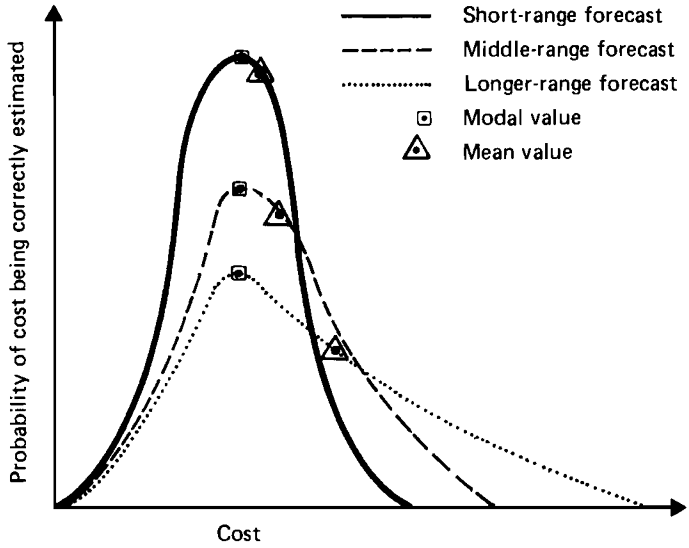


FIGURE 2 Schematic representation of the Bayesian curves showing the probability of the recurrent cost being correctly estimated, as a function of predicted cost, for three different forecast intervals.

constant-value cost equal to the present cost, the mean creeps outward. And it is the mean – the "expected value" – that a realistic estimator must use.

This leads to the qualitative conclusion that uncertainty in forecasting leads to an escalation of expected cost with time, as well as an increased discount rate, i.e., future payments are likely to be larger than they are now, for various reasons unknown.

Summarizing, higher inflation leads to a higher investment risk which is reflected in the charge rate on capital. This is characterized by an elasticity factor, λ; for an inflation rate L, the investor will demand an increased real rate of return λL. The value of λ is not likely to be greater than 0.2. The augmented rate of return produced by this elasticity will also be reflected in the discount rate applicable to future recurrent costs.

Uncertainty in non-monetary conditions affecting the change in costs over time tends to increase the expected values of future expenditure. The rate of escalation is highly dependent on the specific expenses being examined, but could well be higher than the increase in the discount rate caused by uncertainties related to inflation. This is the scenario-dependent escalation rate, σ.

Mathematically, this model can be compared with the previous one by setting up a table similar to Table 8. An inflation rate of 6% and levelized-cost accounting are assumed. A value of the elasticity factor $\lambda = 0.2$ is adopted to test the impact of a large uncertainty in the rate of inflation. Table 10 gives the discount and charge rates obtained.

TABLE 10 Components of the prices charged by utilities, taking into account the uncertainty due to inflation. (35-year amortization with 6% inflation, $\lambda = 0.2$.)

Component	Levelized for 6% inflation	Description
Capital cost	$F_0 + L + \lambda L = 10.6375\%$	Discount rate for capital expenses
Interest	$I_0 + L + \lambda L = 9.95\%$	Discount rate for recurring expenses
Amortization	$A = 0.2633\%$	
Taxes	$0.55(F_0 + L + \lambda L) = 6.16\%$	
Capital + Amortization + Taxes	17.0608%	Capital charge rate against price

TABLE 11 Comparison of the prices charged by various utilities, allowing for the uncertainty due to inflation, and using different values for the scenario-induced escalation of recurrent costs (σ). All values in mills/kWh. (35-year levelizing, 9.95% discount – see Table 10.)

Plant type	Capital charge on price	$\sigma = 0\%$ per year		$\sigma = 1\%$ per year		$\sigma = 2\%$ per year		$\sigma = 3\%$ per year	
		Recurrent cost	Total price	Recurrent cost	Total price	Recurrent cost	Total price	Recurrent cost	Total price
LWR	24.4	17.5	41.9	20.3	44.7	23.7	48.1	28.1	52.5
LMFBR	29.2	9.7	38.9	11.3	40.5	13.2	42.4	15.6	44.8
CS	16.5	43.4	59.9	50.3	66.8	58.9	75.4	69.7	86.2
CFB	19.5	34.5	54.0	40.1	59.6	46.9	66.4	55.5	75.0
SS	57.6	2.1	59.7	2.5	60.1	2.9	60.5	3.4	61.0

Table 11 shows prices calculated from the data of Table 10. Levelized accounting over 35 years is assumed. The scenario-induced rate of escalation of recurrent costs is varied in the range 0–3%. Assuming 6% inflation, this means that recurrent costs increase at a rate between 6 and 9%, with a discount rate of 9.95%.

Comparing Table 11 with Table 9, it can be seen that at low "scenario escalation" rates, (0–1% per year) the qualitative assessment of the various technologies remains essentially unchanged. At "scenario escalation" rates of 2%, and even more at 3%, technologies with low recurrent costs, LMFBR and solar, improve their relative economic ranking. By and large, however, the results of the model confirm the more naive conclusions given in Table 9.

There is one other way of dealing with uncertainty: shorten the levelizing period. This is often done simplistically. The amortization charge is varied while retaining the rest of the capital charge structure of Table 10. For example, if a 15-year amortization period is taken, the amortization charge of Table 10 is increased to 2.7058%, and the total capital charge rate against price rises to 19.5033%. The levelizing period for recurrent expenses is also reduced to 15 years.

There are many ways of tackling this problem, and the results of a number of approaches are presented in Table 12. The first column of levelized costs refers to the capital charges of Table 10, adjusted as above but with no scenario escalation. In the next column, the present worth of the plant after 15 years, computed from Eq. (3), is subtracted from the original capital cost and the calculation repeated. After 15 years, the present worth of a 35-year (real) amortizing plant is 18% of its original value. The third column of levelized costs carries out the same calculation using Eq. (4), and under these conditions the present worth of the plant after 15 years is shown to be 42% of its original value. Finally, the last two columns list the levelized recurrent costs of the subsequent 20 years of operation, expressed both in dollars of the year of commissioning (1978) and in dollars current at the end of 15 years (1993).

How are these numbers to be interpreted? The data in the first column of costs can be dismissed as being exceptionally naive. They are the result of taking a tax adjustment (the use of a fictitious amortization time) literally, and assuming that the plant has no capital value thereafter. Note that this is the only column which places the price of coal-derived electricity within 20% of that of nuclear power, under the cost figures used in these examples. The method used to obtain the figures in the next column has the virtue of recognizing that this short write-off period does not take into account the value of the plant at the end of that period. However, for reasons discussed previously, the use of a high discount rate underestimates this value, i.e., the sale value of a 15-year-old plant 15 years from now will probably be more than 35% of the original capital cost in constant-value dollars, a number consistent with a present worth of 18%. The data in the third column of levelized costs have been corrected to give the plant a significantly higher present worth 15 years in the future, but may indeed have overcompensated, in spite of the arguments used above. A sale price 82% of the original capital cost, in constant dollars, is predicted 15 years in the future.

The best values to use for the levelized cost after 15 years are likely to lie between the values in these two columns, and could only be evaluated more precisely by estimating future values in detail. Finally, the last two columns also refer to the future value of the plants: the lower the recurring costs over years 15–35, the greater will be the incentive to use the plant. These figures demonstrate the great advantages of nuclear power, in particular the breeder reactor, and the great potential of solar power.

The fourth column of Table 11 (total price, 35-year levelizing, no scenario escalation) can be obtained by multiplying the second column of Table 12 (15-year write-off, no capital value thereafter) by 0.875 and adding to this the fifth column of Table 12 (levelized recurring cost over years 15–35 in 1978 dollars) multiplied by 0.511. These coefficients indicate that the operating costs of the system over years 15–35 are not insignificant in determining the long-term value of the plant.

TABLE 12 Levelized costs calculated after 15 years under various assumptions, and levelized recurrent costs over years 15–35.

	Levelized costs after 15 years			Levelized recurrent costs over years 15–35	
Plant type	15-year write-off	Corrected for plant value using F	Corrected for plant value using F_0	1978 $	(6% inflation) 1993 $
LWR	40.3	37.5	33.8	13.0	31.2
LMFBR	40.2	36.9	32.8	7.2	17.4
CS	49.6	47.7	45.2	32.4	77.5
CFB	46.7	44.5	41.5	25.8	61.7
SS	67.3	60.8	52.1	1.6	3.9

12 DISCUSSION

This research was originally motivated by the discrepancy between two cost ratios: the ratio of present worth of future expenses to capital costs; and the ratio of operating and fueling expenses to capital charges, as presented in many discussions on the cost of electrical power. It quickly became clear that all consistent, standard accounting methods (of which the present-worth technique is one) would give the same answers when comparing the costs of various plants. However, systems that combine capital charge rates measured in current dollars with the expenses accrued over the first year or first few years grossly underestimate the contribution of recurring costs to the actual cost during an inflationary period. The effect of this "mixed-mode" accounting is still felt, albeit at a lower level, when prices, rather than costs, are compared.

The recurring costs of fuel and labor are a much larger proportion of the cost of providing electrical power than one is often led to believe. For fossil fuels, *including coal*, these costs are so high that it would take a major collapse of their price structure, or a drastic increase in the relative cost of nuclear plants to make coal-fired systems competitive with nuclear power, i.e., with LWRs as they exist today. Further, looking ahead to future developments, those systems which minimize recurring costs will have a significant advantage over the others. If a breeder reactor (LMFBR) could be provided at twice or three times the cost of a coal-fired plant, and if its target costs for fuel cycle operations are achieved, the breeder reactor immediately becomes the reference (cheapest) source of electrical power. If a solar-electric plant with sufficient energy storage for base-load use could be built at a cost only about four times that of a coal-fired plant, it would be competitive. These capital cost targets are much less forbidding than the goals often cited: factors of 1.25 for LMFBR over LWR, 3 or less for solar power over coal. Indeed, many would argue that a capital cost target for LMFBR twice that of LWR is already within our grasp. (However, it is possible that even the relatively low target suggested here for solar power may not be achieved.)

The same reasoning also suggests that other nuclear electrical-energy generating systems might be more economical than LWRs. One example is the CANDU reactor, which is now being used in Canada. The capital costs of this system are probably less than 50% higher than those of LWRs, when first cores (more expensive for the LWR reactor) and heavy water (for the CANDU reactor) are included in the capital cost. The recurring costs of the CANDU reactor, which requires less uranium, little or no enrichment, and less expensive fuel fabrication, could well be less than half those of an LWR. If these rough estimates can be verified by more careful engineering evaluations, the CANDU reactor could be a suitable power system for the United States today.

In an attempt to reduce the effect of uncertainties, evaluations are sometimes based on projections of the cost for the first few years of operation only. This is an approximation to mixed-mode accounting, particularly when levelized costs in current dollars are being projected. Heuristically, this method can be criticized for ignoring the physical and economic value of the plant beyond the levelizing period. It favors technologies with high recurrent costs even though it is precisely these technologies whose long-term costs are the most uncertain.

In a time of high and uncertain inflation, a utility finds it reassuring to use current-dollar accounting to recover capital investments. Since the present worth of each year's payment decreases rapidly with time, the capital is recovered quickly. However, this does not relieve the planner of his obligation to estimate recurrent costs over the entire plant lifetime. Indeed, the very fact that when the costs of a number of systems are compared after long and short levelizing periods the results are different, shows that great care must be taken in assessing the economic values of the plants at various stages in their lifetimes.

One conceptual flaw in utilizing current-dollar accounting during a period of inflation is the question of discontinuity. Both current-dollar capital payments and levelized recurrent payments generate excess income early in system operation and, in terms of constant-value dollars, the long-term future is subsidized by this excess. Existing plants in a utility system then *seem* to be producing power much more cheaply than is possible for any new plant. The introduction of a new plant is then always seen by the consumer as a diseconomy. This accounting method requires that each application for a new plant be accompanied by an application for a rate increase. Constant-dollar accounting avoids this unpopular measure.

It is sometimes alleged that fuel escalation pass-through allowances (i.e., letting the price paid by the consumer rise to cover the inflating price of fuel) are a prime reason for utilities to prefer high-recurrent-cost fossil-fuel technologies. However, this does not seem to be tenable within the logic of the industry. While pass-through allowances protect the utility against out-of-pocket losses, they also increase the consumer price and inhibit the use and growth of the utility system. The practice of giving pass-through allowances serves to consolidate the position of mixed-mode accounting in the power-generating industry, and any effect this may have on utility planning is a function of the method of accounting employed.

The methods of accounting presented in this report are not new, and have only been presented to illustrate that the basic principles of elementary engineering economics seem to have been violated routinely in utility planning. Neither are the detailed results particularly new. Stauffer *et al.* (1975a,b) have examined the case for the breeder reactor using the accounting methods discussed above (including full levelizing of recurrent costs over 30 years) and came to the same conclusions reached here with regard to economic targets. They also found the comparative costs of coal plants to be high. It is interesting that, despite the intervening period of inflation, these papers, presented in 1975, using quite different input numbers – essentially the cost of plants, fuels, and fuel cycle operations prevalent in 1974 – reached the same conclusions found today. The present report goes further in that it includes solar power in the comparisons, updates the input, and examines the discrepancy between consistent planning results and operational decisions.

Two factors have been omitted from this discussion, and should be explicitly included in any detailed planning operation:

1. It has not been normal practice to collect capital costs in constant-value dollars as these costs are accrued, nor to inflate past expenditure (anti-discount) to dollars of the commissioning year. Adhering to correct practice, could, under present inflationary conditions, add of the order of 20% to the *real* capital cost of most of the plants examined. Plants with high capital costs will therefore suffer in comparison with plants whose capital costs are lower.
2. Any planning operation must include a projection for the capacity of each plant considered. Because of their high operating costs, fossil-fueled plants will be run at a lower level as they grow older. This penalizes them in comparison with other types of plant. It seems almost mandatory that "discounted" capacity factors be calculated in constant-dollar formulations, for the reason discussed with regard to amortization. Otherwise, as with mixed-mode accounting, the long-term economic value is lost. Again, proper accounting improves the comparative rating of systems with low recurrent costs.

The logic behind the regulatory control of utilities has been touched on only superficially. This control includes not only the regulation of prices, but also the socio-economic controls implicit in taxation, licencing, financing requirements and rules, and the granting of franchises. These are regarded as external to the planning of the utility, and will be discussed in a later paper, which will consider social profit and loss.

ACKNOWLEDGMENTS

Constructive comments received from Harvey Brooks and Ludwig F. Lischer are acknowledged with thanks.

REFERENCES

Corey, G.R. (1977) Testimony to the US House of Representatives, Committee on Government Operations, Subcommittee on the Environment, Energy, and Natural Resources, September 19.

National Research Council (in preparation) Report of Committee on Nuclear Alternative Energy Systems (CONAES). Washington, DC: National Academy of Sciences.

SRI International (1977) Fuel and Energy Price Forecasts: Quantities and Long Term Marginal Prices. Carried out for the Electric Power Research Institute, Report EPRI EA-433.

Statistical Abstract of the United States (1976) United States Department of Commerce, Bureau of the Census.

Stauffer, T.R., H.L. Wyckoff, and R.S. Palmer (1975a) The Liquid Metal Fast Breeder Reactor. Assessment of Economic Incentives. Presented to the Breeder Reactor Corporation, Chicago.

Stauffer, T.R., R.S. Palmer, and H.L. Wyckoff (1975b) Breeder Reactor Economics. Prepared for the Breeder Reactor Corporation. Sonnyvale, California: Fast Breeder Reactor Department, General Electric Company.

A COMPARATIVE STUDY OF PUBLIC BELIEFS ABOUT FIVE ENERGY SYSTEMS

Kerry Thomas, Dagmar Maurer, Martin Fishbein,
Harry J. Otway, Ron Hinkle, and David Simpson

PREFACE

The risks associated with alternative energy systems, and public perceptions of these risks, have become important constraints in the selection of energy strategies. This Research Report presents results of an application of an attitude-measurement methodology which explores the beliefs held by the public with respect to five alternative energy sources. Emphasis is given to a differential analysis of the belief systems of those subgroups most in favor of (PRO) and most against (CON) the use of nuclear energy. Results specific to public attitudes toward the use of nuclear energy have been published (Otway and Fishbein 1977) and an earlier pilot study on this same topic was reported (Otway and Fishbein 1976). An analysis of the determinants of voting behavior in a public referendum on nuclear energy has also been presented (Bowman *et al.* 1978).

This report is based on work of the Joint IAEA/IIASA Risk Assessment Project, and thus it represents a collaboration between the International Atomic Energy Agency and the Energy Systems Program at the International Institute for Applied Systems Analysis.

SUMMARY

Public acceptance is becoming an increasingly important constraint to be taken into account by those responsible for technological policies. Acceptance by the public will depend on their relevant attitudes toward a given technology, and these attitudes will be a function of beliefs about the attributes and probable consequences of the technology in question. This study explores belief systems with respect to five energy sources: nuclear, coal, oil, hydro, and solar. The method used permits comparisons of attitudes and also of the underlying belief dimensions which characterize each energy source.

Two hundred and twenty-four members of the Austrian public took part in this questionnaire survey; the sample was stratified by age, education, sex, and geographical location (Vienna, provincial capital, and rural).

An overall measure of attitude toward each energy source showed that only in the case of nuclear energy was the sample polarized to any degree. For the fossil fuels there was a large measure of moderate favorability, and for the renewable sources virtually everyone expressed a highly favorable attitude.

The major part of the research was concerned not with the overall attitudes of the public but rather with their belief systems, that is with their perceptions of the qualities and attributes of each energy source. A set of 39 attributes of energy sources was used. These attributes were associated in propositional form with each of the five energy sources (e.g., the use of oil leads to water pollution) and the respondents rated their degree of belief/disbelief in each statement.

The data were simplified using factor analysis. Five underlying dimensions of belief were identified as common to all the energy sources. These dimensions were concerned with: future-oriented and political risks; economic benefits; environmental risks; psychological and physical risks; and future technological development. The attributes most clearly identified with each of these dimensions were used, for each energy source, to construct the profiles of beliefs held by the sample as a whole.

The Austrian sample as a whole believed that environmental risks were associated with oil, coal, and nuclear energy, in that order; they believed that all the sources except coal provided approximately the same, moderate level of economic benefit; and that only nuclear energy and solar energy would lead to technological development. The sample believed that only nuclear energy would lead to psychological and physical risks; and they believed strongly that, with the single exception of nuclear energy, none of the sources would lead to indirect (future-oriented and political) risks.

Since nuclear energy was the only case where the attitude measures showed groups in the public both in favor of (PRO) and against (CON) the energy source, belief profiles were constructed for two subgroups – those most and least

favorable toward the use of nuclear energy. When these belief profiles were examined it was clear that treating the sample as a whole masked important information. First, the two groups had very different belief systems about nuclear energy; and second, the two groups had similar perceptions of hydro, solar energy, and coal, although their beliefs about oil were slightly different.

The sample as a whole (even those most favorable toward nuclear energy) preferred the use of hydro and solar energy. This is because both PRO and CON groups saw these two energy sources as less of a threat than nuclear energy on all risk-related dimensions. The PRO group perceived nuclear energy as the source most likely to lead to economic benefits and future technological developments; the lower ratings given to the fossil fuels by this group were primarily due to beliefs that these sources would provide only small economic benefits while leading to appreciable environmental risks. However, the CON group viewed nuclear energy as only marginally more likely than the fossil fuels to lead to economic and technological benefits but as an appreciably greater threat on the risk-related dimensions.

1 INTRODUCTION

Public acceptance is becoming an increasingly important constraint to be considered by those responsible for technological policies. In order to formulate policy wisely it is necessary to understand the underlying determinants, i.e., belief systems, of acceptance or opposition by public groups; in our research we have used the attitude concept for this purpose. The particular approach adopted, in addition to providing an overall estimate of attitude, permits a detailed examination of underlying beliefs. It thus provides a method for exploring systematic differences in belief systems between groups of particular social, political, or professional significance.

The first report in this series (Otway and Fishbein 1976) was a pilot study of the beliefs and attitudes held by a group of energy experts with respect to nuclear energy. This was followed by a similar analysis for a heterogeneous sample of the Austrian public (Otway and Fishbein 1977).[1] The present report describes results of the latter study which extend the exploration of belief systems to include five energy sources: nuclear, coal, oil, hydro, and solar. The beliefs about these five sources held by the entire Austrian sample are described, and a comparison is made between the beliefs held about all energy systems by those subgroups shown to be most in favor of (PRO) and those most against (CON) the use of nuclear energy.

2 METHODOLOGICAL APPROACH

The attitude model used in our studies of the determinants of public acceptance of energy systems has been described in some detail in the reports cited earlier. Therefore we will simply summarize the main points which are relevant to the procedures and analyses discussed in this report.

First, attitude is defined as an overall feeling of favorableness toward an object, where "object" refers to any discriminable aspect of the individual's world. Attitude can be measured either directly, using the semantic differential technique of Osgood *et al.* (1957), or indirectly by considering the responses to a set of belief or opinion items about the attitude object. Second, the model used specifies the relation between beliefs and overall attitude, as follows:

> Each belief is treated as a subjective probability judgment that the attitude object is associated with a given characteristic or attribute. The evaluation of each attribute is then weighted by the probability of the association (i.e., the belief strengths). Thus, according to the model, attitude is approximated by the pairwise products of belief strength × evaluation summed over a set of suitable beliefs.[2]

Strictly, if one wishes to relate beliefs (or observed differences in beliefs between groups) to attitude in a deterministic sense, it is necessary to use only salient beliefs. These are the beliefs which are within the span of attention of each individual when the attitude is measured. In most practical situations, however, a set of modal salient beliefs is used, i.e., those beliefs occurring most frequently in the sample.

In this study a set of modal beliefs about the attributes of energy sources was chosen on the basis of interviews with members of the general public, the data collected in previous research, and a literature survey. The complete set of 39 attributes (see Table 3) spans the most commonly perceived, possible consequences of using coal, oil, hydro, solar, and nuclear energy. Since the initial concern was with perceptions of nuclear energy, some of the items are specific to this particular source. It follows that, as a set, the 39 belief items cannot be interpreted as "salient" (using Fishbein's terminology) for each and every energy source. Therefore it would be incorrect to make generalizations about the contributions of these beliefs to attitudes toward all energy sources. This report therefore focuses on *strength of belief* data, that is, on the public's beliefs and perceptions of the energy sources, without any necessary implication for the determination of specific attitudes. There is one exception to this: in the case of nuclear energy the same set of 39 attributes has been successfully used in the same attitude model to explore the public acceptance of nuclear energy (Otway *et al.* 1978). The purpose of the present paper is to examine how attributes, already shown in the earlier study to contribute to attitudes toward nuclear energy, are perceived by the public in relation to other energy sources. Particular attention is given to contrasting perceptions of coal, oil, hydro, and solar energy held by those subgroups of the general public who are most in favor of (PRO) and most against (CON) the use of nuclear energy.

3 METHOD

SAMPLE

Sampling of the general public was not intended to be representative of the Austrian population but was a stratified sample controlling for geographic location (Vienna, provincial capital, and rural), sex, age, and education. The total number of usable interviews was 224* and the breakdown of this total across the demographic categories is shown in Table 1.

QUESTIONNAIRE

Apart from demographic information the questionnaire measured the following three factors: overall attitude toward each energy system, attitudes toward each of the 39 attributes (attribute evaluation), and belief strengths.

Overall Attitude toward Each Energy System

This was measured using the semantic differential technique of Osgood *et al.* (1957), i.e., the rating of each attitude object on a series of 7-point scales (+3 to −3) with the end-points labeled with adjective pairs such as good/bad, harmful/beneficial. In keeping with Osgood's procedure, a factor analysis of the responses to these scales, for all five energy sources, was used to identify adjective pairs which most clearly represented the evaluative dimension, which is the dimension that Osgood has equated with attitude. Five adjective pairs were validated in this way and used in the remaining analyses: good/bad, harmful/ beneficial, harmonious/controversial, acceptable/unacceptable, moral/ immoral. The measure of overall attitude was a sum of the ratings on these five scales giving a range of +15 to −15.

*However, in a small number of cases, respondents did not *completely* fill in the questionnaire: it will therefore be noticed that the sample size for particular sections is sometimes less than 224.

TABLE 1 Demographic breakdown of the Austrian public sample ($N = 224$).

		Vienna ($N = 121$)		Provincial capital ($N = 51$)		Rural area ($N = 52$)		
Education level	Age	Male ($N = 81$)	Female ($N = 40$)	Male ($N = 29$)	Female ($N = 22$)	Male ($N = 31$)	Female ($N = 21$)	ΣN
Grade school	18–34	6	3	2	1	1	5	18
($N = 45$)	35–50	4	1	3	3	2	3	16
	51–65	2	2	1	1	4	1	11
Trade school	18–34	11	14	5	7	8	4	49
($N = 80$)	35–50	3	2	3	6	2	1	17
	51–65	6	2	2	1	2	1	14
High school/university	18–34	30	9	8	1	7	4	59
($N = 99$)	35–50	12	3	4	1	3	1	24
	51–65	7	4	1	1	2	1	16

Attitudes toward Each of the 39 Attributes (Attribute Evaluations)

These were measured in a similar fashion but using only a single 7-point scale (+3 to −3) labeled with the adjective pair good/bad. Each attribute was presented without reference to any specific energy source. For example,

Increasing the prestige of my nation

GOOD :–:–:–:–:–:–:–: BAD

Belief Strengths

These were measured by relating the 39 attributes to each energy source in turn and asking the subject to indicate his judgment of the truth of the statement. A 7-point scale (+3 to −3) was used and the end points were labeled likely/unlikely. For example,

The use of coal leads to air pollution

LIKELY :–:–:–:–:–:–:–: UNLIKELY

It should be noted that although belief strength has been construed as a subjective probability, the way it is scaled (in keeping with most of Fishbein's own work) avoids certain strict requirements of probability measures. The beliefs are *not* treated as a partitioned event space where the probabilities would sum to 1, and further, by using the bipolar scale (+3 to −3) it is possible to encompass levels of probability that the energy source *is* or *is not* associated with the attribute in question.

4 RESULTS

Although the primary concern of this report is the comparison of beliefs about using different energy sources, it is worthwhile to consider first the overall feelings, or attitudes, toward the different sources of energy generation.

ATTITUDES TOWARD FIVE ENERGY SOURCES

Examination of the attitude scores in the total sample (as measured by the semantic differential) yielded the three distinct types of frequency distribution shown (smoothed) in Figure 1. The distributions were virtually the same for the two fossil fuels, as were those for hydro and solar energy; however, the distribution for nuclear energy was quite different. In the case of fossil fuels there were very few negative attitudes and few highly positive; most respondents were moderately favorable. For hydro and solar energy there were virtually no negative attitudes; the most frequent response was highly favorable. Attitudes toward nuclear energy centered in the middle of the scale but with clusters of highly negative and highly positive attitudes at both ends. It was only in the case of nuclear energy that attitudes were sufficiently polarized to warrant differential analyses of underlying beliefs for "PRO" and "CON" groups.

As in the earlier study, two subgroups were formed from the total sample by selecting the 50 respondents most favorable to the use of nuclear energy (PRO group) and the 50 most against its use (CON group). Differences in attitude held by the PRO and CON groups toward the remaining four energy sources were examined by analysis of variance (ANOVA).

The mean values of attitude for each group with respect to energy sources are shown in Table 2. In general, the PRO nuclear group was more favorable toward the non-nuclear energy sources (mean = 10.6) than was the CON nuclear group (mean = 7.9). There was a main effect of energy source on attitude scores, i.e., significant differences in attitudes toward the different sources were observed. For the total sample, respondents were generally more favorable toward

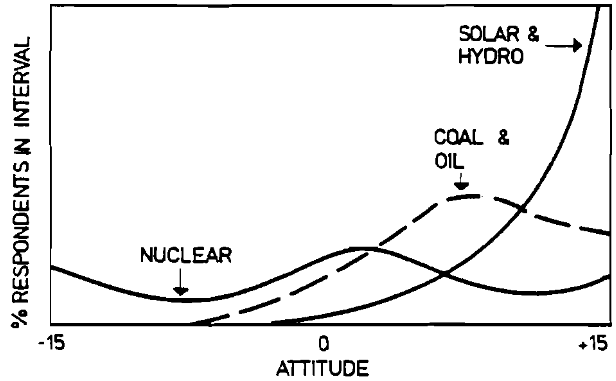


FIGURE 1 Smoothed frequency distribution of attitudes toward energy sources.

hydro (10.7) and solar energy (10.5) than they were toward coal (6.1) and oil (5.4); they were least favorable toward nuclear energy (0.4).

There was also a significant interaction effect which, in this case, indicated that those PRO and CON nuclear energy had similar attitudes toward hydro and solar energy, but differed in their attitudes toward each of the two fossil fuels. The largest difference between the PRO and CON nuclear groups (apart from their attitude to nuclear energy) was their attitude toward oil as a source of energy, the PRO group being significantly more favorable toward its use. When comparisons were made (within the PRO and CON groups) between attitudes toward each possible pair of the four non-nuclear energy sources, those PRO nuclear energy had significantly different attitudes toward all pairs except solar/hydro and coal/oil. The CON group had different attitudes toward all possible pairs except solar/hydro.

To summarize briefly, the PRO nuclear group was more favorable to hydro and solar energy than to coal and oil. Their attitudes toward nuclear energy did not differ appreciably from their attitudes toward oil, and their attitudes toward both nuclear and oil were significantly less favorable than those toward hydro and solar and somewhat more favorable than those toward coal. In contrast, the CON nuclear group was strongly negative toward nuclear energy but had positive attitudes toward the other energy sources; they were most favorable toward hydro and solar, moderately favorable toward coal, and significantly less favorable toward oil.

UNDERLYING COMMON DIMENSIONS OF BELIEF ABOUT ENERGY SOURCES

In the earlier report on beliefs and attitudes of the public toward the use of nuclear energy (Otway and Fishbein 1977) it was found, using factor analysis of belief-strength scores, that the 39 beliefs about nuclear energy clustered on

TABLE 2A Mean values of attitudes of those PRO and CON nuclear energy toward five energy sources.

Group	Energy source					
	Nuclear	Solar	Hydro	Coal	Oil	All[a]
PRO ($N = 50$)	(10.2)	12.2	12.3	8.3	9.7	10.6
CON ($N = 50$)	(−10.1)	11.1	11.2	6.2	3.1	7.9
	**	NS	NS	*	**	**
Total sample ($N = 218$)	(0.4)	10.5	10.7	6.1	5.4	8.2

*Difference between groups significant, $p < 0.05$.
**Difference between groups significant, $p < 0.01$.
NS, difference between groups not significant.
[a]All refers to all energy sources *except* nuclear.

TABLE 2B Summary of analysis of variance of attitude toward five energy sources held by those PRO and CON the use of nuclear energy.

Main effects	
PRO/CON (*A*)	$p < 0.001$
Energy sources (*B*)	$p < 0.001$
Interaction	
$A \times B$	$p < 0.001$

four factors.[3] These dimensions underlying perceptions of nuclear energy were named psychological risk, economic/technical benefits, sociopolitical risk, and environmental/physical risk. The reduction of the belief set to four major dimensions, in practical terms, facilitated comparisons between those who were PRO and CON nuclear energy. In order to identify commonalities in perceptions of the five energy sources it again seemed reasonable to reduce the set of 39 items to a smaller set of underlying dimensions by using factor analysis. In this case Tucker's (1966) extension of the factor-analytic procedure to three-dimensional matrices ($n \times m \times q$, where n subjects responded to m belief statements about q energy sources) were used.[4] The three modes in this analysis were thus

- The source mode, five energy sources
- The belief mode, 39 attributes of energy sources
- The subject mode, 224 members of the Austrian public

The findings are reported briefly for each of the three modes in turn, followed by a detailed analysis of the belief mode.

Energy Source Mode

The three-mode factor analysis identified three source factors, one for nuclear energy, one for the fossil fuels, and one for hydro and solar energy. This finding is consistent with the frequency distributions of attitude scores which showed one pattern for the fossil fuels, another for hydro and solar energy, and a different distribution for nuclear energy.

Belief Mode

It will be recalled that the earlier report, based on the Austrian public's beliefs about nuclear energy, showed that four underlying dimensions could account for the intercorrelations amongst the 39 beliefs (i.e., psychological risks, economic/technical benefits, sociopolitical implications, and environmental/physical risk). When three-mode factor analysis was used to identify commonalities amongst perceptions of all five energy sources, the best solution changed slightly and five factors emerged.

The factor structure for beliefs about all energy sources differed from that for nuclear energy alone primarily in that, when the five sources were considered together, the economic/technical benefits factor separated into two factors: an *Economic Benefits* factor, and a future-oriented *Technology Development* factor. In addition, the psychological risk factor associated with nuclear energy included physical risks when all five sources were considered (*Psychological and Physical Risk* factor). The sociopolitical factor associated with nuclear energy became a more general, future-oriented, and political factor which is now called *Future and Political (or Indirect) Risk*. The fifth dimension remained an *Environmental Risk* factor. The five attributes most closely associated with each of these five factors are listed in Table 3.

Subject Mode

Three subject factors were found. Subject Factor I was related to the subjects' strength of agreement with the modal view of the energy sources. Those high on Factor I tended to respond in the same direction (be it positive or negative) as the sample mean, but more extremely; those low on Factor I also tended to respond in the same direction, but less extremely than the sample mean. Thus, in the context of substantial agreement as to the direction of relationships between the energy sources and various attributes, the subjects' strength of belief was a function of their Factor I scores. This factor may be simply a response style, or a tendency to use the ends of the response scale. However, supplementary analyses of Factor I scores, as a function of demographic variables, suggest

TABLE 3 The belief dimensions and most characteristic belief items identified by three-mode factor analysis.

Belief dimension	Belief item
Economic benefit	Good economic value Increased standard of living Increased employment The industrial way of life Increasing Austrian economic development
Environmental risk	Air pollution Water pollution Production of noxious waste Making Austria dependent on other countries Exhausting our natural resources
Indirect risk (Future-oriented and political)	Changes in man's genetic make-up Increasing rate of mortality (not) A technology I can understand Formation of extremist groups A police state
Technological development	New forms of industrial development New methods in medical treatment Dependency on small groups of experts Technical spin-offs (not) Exhausting natural resources
Psychological and physical risk	Accidents which affect large numbers of people Exposure to risk which I cannot control Rigorous physical security measures Hazards caused by human failure Hazards caused by material failure
Belief items not strongly identified with the five belief dimensions	Exposure to risk without my consent A threat to mankind Risky Delayed effect on health Increases my nation's prestige Reduces the need to conserve energy Satisfies the energy need in the decades ahead Decreases dependence on fossil fuels Increases the extent of consumer orientation Diffusion of knowledge about construction of weapons Transporting dangerous substances Destructive misuse of technology by terrorists Gives political power to big industrial enterprises Increases occupational accidents Long-term modification of the climate

that this tendency to make more extreme responses may be interpreted as greater confidence, and may, in fact, reflect greater knowledge. Specifically, individuals' scores on this factor were positively related to age and education, and to prestige as based on measures of socioeconomic status and occupation. Further, males scored significantly higher on this factor than did females. The extent to which an individual was identified with this "confidence" factor did not correlate significantly with attitude toward nuclear energy (r = 0.02), but correlated positively with attitudes toward hydro (r = 0.40) and solar energy (r = 0.43). The correlations with attitudes toward the fossil fuels were also significant but low (r = 0.29 and 0.27, for coal and oil, respectively).

Subject Factor II was more obviously a response style mode; those scoring high on this factor were invariably closer to the "unlikely" or negative side of the scale, regardless of the content of the item or the implication of the scaling response. Scores on this factor were not significantly correlated with attitudes toward any of the five energy sources. Of the demographic variables, only age showed a significant relationship with Factor II scores. The 24–34 age group had high scores on Factor II while the scores of all other groups (under 24, 35–50, and over 50) were low. Thus, age group 24–34 had a tendency to see all relationships between energy sources and attributes as relatively less likely. This finding for some of the younger participants could be interpreted as a general "negativism," or it could indicate that the attributes used in this survey were less relevant for the 24–34 age group than for the rest of the sample.

Subject Factor III appeared to be a "true" content dimension. Those subjects who had low scores on Factor III shared three common viewpoints:

- They perceived all five energy sources as economically viable, a perception not shared by the modal view (note that the group as a whole, for example, saw coal as an uneconomic prospect)
- They saw nuclear energy as generally "better" than the modal perception, being, for example, more likely to be economically sound and to lead to technological (spin-off) developments
- They perceived oil as somewhat better on all counts than the modal view, being, for example, less likely to lead to indirect risks and more likely to lead to technological spin-offs

This summary of the viewpoint of those individuals who scored low on Factor III (diametrically opposing views were held by those with high scores on Factor III) shows that this subject factor represents an underlying dimension which primarily relates to beliefs about nuclear energy. Consistent with this explanation it was found that Factor III scores correlated with the semantic differential measure of attitude toward nuclear energy (r = –0.59). Factor III scores also correlated with attitudes toward the fossil fuels (r = –0.42 and –0.23, for oil and coal, respectively). Of the demographic variables, only age showed a significant relationship to Factor III scores. The 24–34 age group had high scores on Factor III, the 35–50 group was relatively neutral, and the scores of the "under 24" and "over 50" groups were low.

In summary, the interpretation of the three-mode factor analysis is straightforward for the energy mode and the belief mode: the sample of the Austrian public perceived nuclear energy differently from other sources, but perceived the two fossil options as similar, and also hydro and solar energy as similar. For the belief mode five factors emerged: psychological/physical risk, economic benefits, technological development, future/political risk, and environmental risk. These dimensions represent the basic considerations that are taken into account in judging the different energy systems. The findings for the subject mode are more difficult to interpret since the "types" which emerged could not be definitively identified by demographic variables (i.e., they were not clearly specified social groups).

The analysis of the subject mode indicated that there were three sorts of considerations that influenced respondents' judgments about the attributes of the five energy systems

- A "confidence" factor where (on many items) the sample is in general agreement that a given energy source has (or does not have) a particular attribute, but some people tend to be more confident (or extreme) than others (Factor I)
- An influence of response style whereby some people tended to use the "unlikely" side of any scale (Factor II)
- A "true" content dimension that reflects differences in beliefs about the different energy systems (Factor III)

This latter content dimension is notable in that it does tend to distinguish between those who are PRO (low scores on Factor III) and CON (high scores on Factor III) nuclear energy. That is, the viewpoint of those individuals scoring low on Factor III was similar to that of the original PRO nuclear group used in our earlier reports.[5] Further examination showed that 56% of the PRO group was present amongst the 50 lowest scores on Factor III, and 52% of the CON group was present amongst the 50 highest Factor III scores. Despite this overlap it is not reasonable to assume that the two groups correspond sufficiently to generalize *a priori* from the Factor III findings to a PRO–CON analysis. However, analysis of variance of beliefs about the five energy sources, based on these two alternative groupings (either low/high scores on Factor III or the original PRO–CON nuclear groups), showed very similar results. While it is of some interest to examine the different belief systems of subjects low and high on Factor III, it must be recalled that respondents' final judgments are influenced not only by their position on Factor III, but also by their positions on Factors I and II. Therefore, in keeping with the earlier reports and with the basic social question underlying the research, the remainder of this report will primarily consider the beliefs of those public groups who were most in favor (PRO) and most against (CON) the use of nuclear energy.

PUBLIC BELIEFS ABOUT FIVE ENERGY SOURCES

The five dimensions underlying perception of the energy options, obtained from the three-mode factor analysis, were used first to examine the beliefs of the Austrian public sample as a whole, and then to compare the belief systems of those PRO and CON nuclear energy. The five belief items most closely identified with each belief dimension were summed to give an index of belief strength ($\Sigma_{i=1}^{5} b_i$) for each energy source in turn. The mean values of $\Sigma_{i=1}^{5} b_i$ for each of the five belief dimensions and each of the five energy sources are shown in bar diagram form in Figure 2 (total sample, $N = 211$). It can be seen that, overall, the public have very different perceptions of the five energy systems. These differences can best be seen by considering each of the five belief dimensions separately.

Indirect Risk

Although the public (on average) believed that none of the five energy sources would lead to future-oriented and political risks (such as a "change in man's genetic makeup" or "a police state"), they were significantly less certain of this vis-à-vis nuclear power than for any other energy source. They were also somewhat less certain that the use of oil would avoid such indirect risks in comparison with coal, hydro, or solar energy.

Economic Benefit

With the exception of coal, the public believed that all energy sources would lead to economic benefits (e.g., "an increased standard of living," or "increased employment"). They believed that oil was the energy source most likely to lead to these benefits, although not significantly more so than hydro or nuclear energy; but all of these three were seen as more likely to lead to economic benefits than was solar energy.

Environmental Risk

Here, on average, the public saw significant differences amongst all the energy sources. They believed that the fossil fuels and nuclear energy would lead to environmental risks (such as air and water pollution) whereas hydro and solar energy would not. The order from most to least risky in environmental terms was: oil, coal, nuclear, hydro, solar; thus the fossil fuels were seen as posing a greater environmental threat than nuclear energy.

Psychological/Physical Risk

Only the use of nuclear energy was perceived as leading to psychological and physical risks (e.g., "accidents affecting large numbers of people," or "exposure

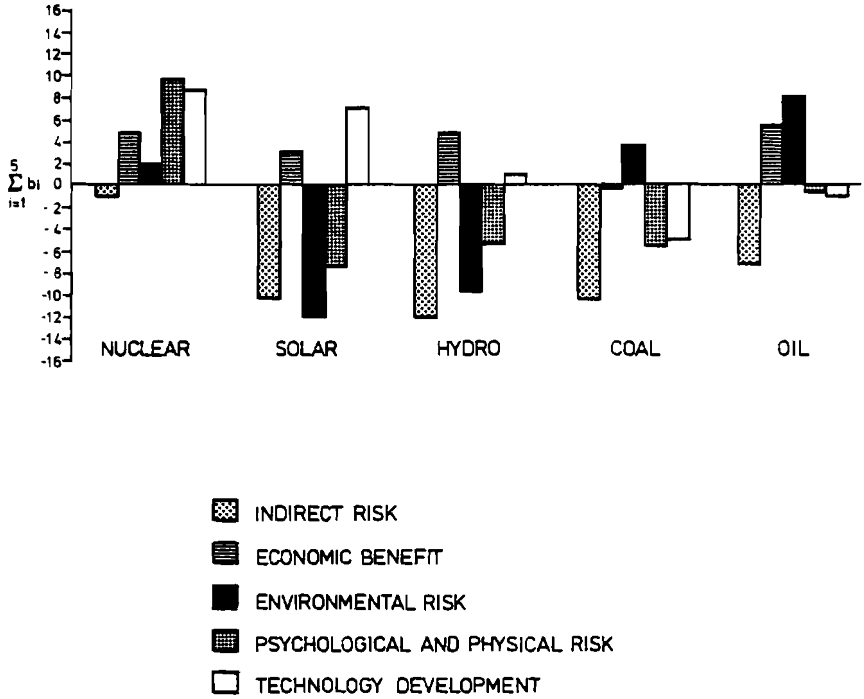


FIGURE 2 Public beliefs about five energy sources ($N = 211$) held by the total public sample.

to risk without personal control"). Solar energy was seen as least risky in this respect, and the public were uncertain with regard to oil.

Technological Development

The public, on average, also saw large differences amongst the energy sources in terms of their likelihood of leading to future technological developments: they were certain that the use of nuclear energy would lead to such developments and that the use of coal would not. They also believed that the use of solar energy would lead to these developments (although statistically less so than nuclear energy), and they were uncertain about oil and hydro in this respect.

DIFFERENTIAL ANALYSIS OF PRO AND CON NUCLEAR GROUPS

While the above results describe the average responses of the total public sample, it is perhaps more meaningful to examine the differing views of the five energy systems which are held by those PRO and CON nuclear energy. These differences were also examined by analysis of variance.[6] As expected, a significant three-way interaction was obtained indicating that, for at least some of the energy sources,

TABLE 4 Mean belief strengths for each belief dimension and energy source held by those PRO and CON the use of nuclear energy.

		Energy source				
Belief dimension	Group	Nuclear	Solar	Hydro	Coal	Oil
Indirect risk	PRO	−6.8	−10.7	−12.2	−10.5	−8.8
(Future-oriented/political)	CON	3.9	−10.5	−12.4	−10.7	−6.6
		**	NS	NS	NS	**
Economic benefits	PRO	7.1	3.9	6.1	1.8	5.5
	CON	0.8	2.6	2.2	−1.6	4.0
		**	NS	**	**	NS
Environmental risk	PRO	−2.7	−11.7	−10.1	3.2	4.7
	CON	5.1	−12.6	−9.9	3.4	9.1
		**	NS	NS	NS	**
Psychological and physical risk	PRO	4.4	−7.6	−6.6	−6.9	−3.5
	CON	12.4	−9.5	−5.9	−5.6	−0.9
		**	NS	NS	NS	*
Technological development	PRO	9.1	5.9	1.7	−5.0	1.3
	CON	6.4	6.5	−1.2	−5.8	−0.8
		*	NS	**	NS	*

*Difference between PRO and CON group significant, $p < 0.05$.
**Difference between PRO and CON group significant, $p < 0.01$.
NS, difference between groups not significant.

those PRO and CON nuclear energy had different beliefs. These differences are given in Table 4 and are summarized in bar diagrams in Figure 3.

It is not surprising that the PRO and CON groups were found to have quite different perceptions of nuclear energy. For the PRO group nuclear energy was believed to lead to economic benefits and technological development, but also to be associated with some degree of psychological and physical hazard. The PRO group did not believe that using nuclear energy would lead to indirect (i.e., future-oriented and political) risks nor, to a lesser degree, to environmental risk. The CON group believed nuclear energy would lead to all three types of risks. They also believed that it would lead to technological developments (but to a lesser degree than did the PRO group), and they did not perceive nuclear energy as leading to economic benefits. The differences between the PRO and CON groups' perceptions of nuclear energy have been discussed in depth in earlier publications (Otway and Fishbein 1977; Otway *et al.* 1978).

Turning to the other energy sources, Table 4 and Figure 1 show that, although those who were PRO and CON nuclear energy did not differ in their beliefs about solar energy, there were significant differences in some of their beliefs about the remaining three energy sources:

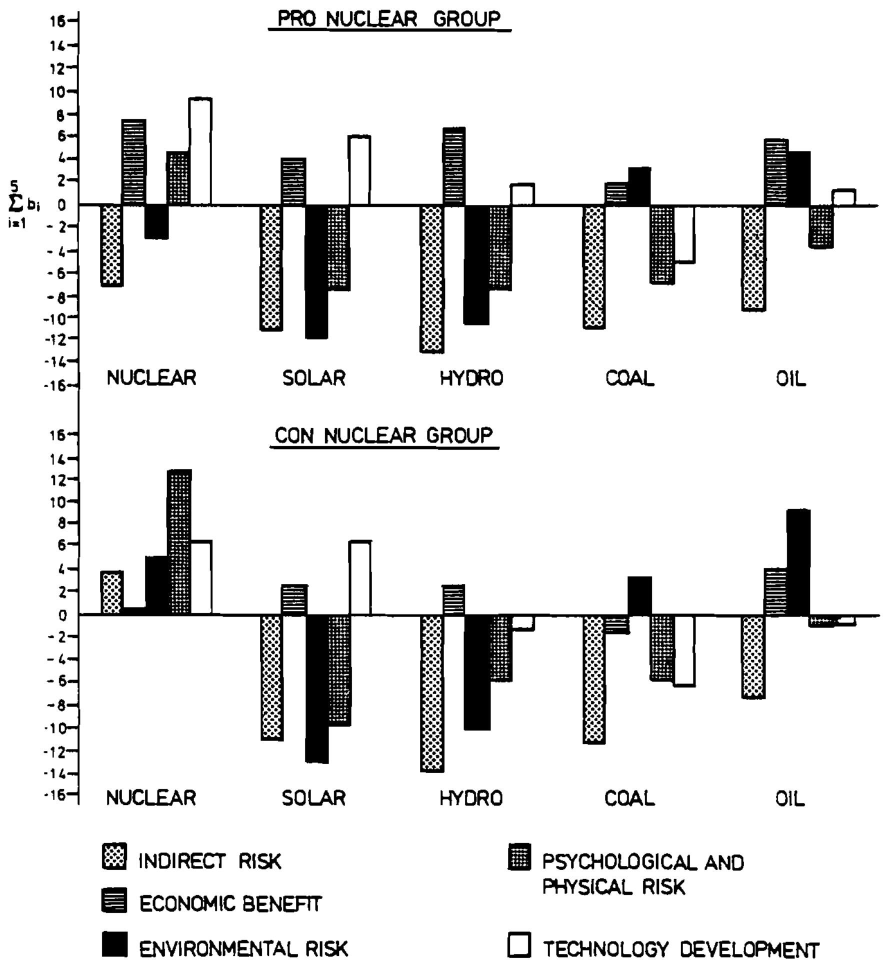


FIGURE 3 Beliefs about five energy sources held by those PRO and CON the use of nuclear energy.

Hydro

On average, people who were PRO or CON nuclear energy believed equally strongly that hydro-power would not lead to any type of risk. They disagreed, however, about the benefits of using these systems. Those who were PRO nuclear energy believed more strongly that their use would lead to economic benefits and technological developments than did the CON nuclear group.

Coal

People who were PRO and CON nuclear energy did not differ in their beliefs about the risks associated with the use of coal, or in their beliefs that using coal would not lead to technological developments. There was a significant difference between the two groups only with respect to economic benefits: the PRO group believed that coal would lead to some economic benefits while the CON group did not.

Oil

The two groups differed more in their beliefs about the use of oil than about any other source apart from nuclear energy; indeed it was only with respect to economic benefits that there was any agreement at all. Consistent with the previous findings that the PRO group's attitude toward oil was more favorable than that of the CON group, the PRO group saw the use of oil as less risky on all counts, and more likely to lead to technological developments.

These different beliefs about the energy sources resulted in different rankings of these sources by the PRO and CON groups. Table 5 shows the differences in mean belief scores, on each dimension, amongst all possible pairs of energy sources. Differences between the PRO and CON groups were found primarily in three areas: comparisons between nuclear energy and the other energy sources, comparisons between hydro and solar energy, and comparisons between coal and oil. These differences will be discussed separately below.

Nuclear Energy As Compared to the Fossil Fuels

Both those groups PRO and CON nuclear energy believed that this energy source was more likely than the fossil fuels to lead to indirect risks as well as psychological/physical risks. However, with respect to environmental risks, nuclear energy was viewed by the PRO group as being less of a threat than the fossil fuels, and by the CON group as being less risky than oil but about the same as coal. Both groups believed that the use of nuclear energy was significantly more likely to lead to technological developments than was the use of either fossil fuel. In terms of economic benefits nuclear energy was seen by the PRO group as a significantly better prospect than coal but only slightly better than oil. In marked contrast, those opposed to nuclear energy believed that oil was the energy source most likely to lead to economic benefits; they saw little difference in this respect between nuclear energy and coal.

Nuclear Energy As Compared to Hydro and Solar Energy

Both PRO and CON nuclear groups believed that hydro and solar energy posed the least threat on all risk dimensions, and significantly less so than nuclear energy. With respect to benefits, however, the PRO group believed that using nuclear energy was significantly more likely to lead to technological developments than either hydro or solar, and likely to lead to significantly more economic benefits

TABLE 5 Pairwise contrasts of belief strengths about different energy sources held by those PRO and CON the use of nuclear energy.

		Indirect risk			Economic benefit			Environmental risk			Psychological/ physical risk			Technological development		
Nuclear/Solar	PRO	−6.8	−10.7	**	7.1	3.9	*	−2.7	−11.7	**	4.4	−7.6	**	9.1	5.9	**
	CON	3.9	−10.5	**	0.8	2.6	NS	5.1	−12.6	**	12.4	−9.5	**	6.4	6.5	NS
Nuclear/Hydro	PRO	−6.8	−12.2	**	7.1	6.1	NS	−2.7	−10.1	**	4.4	−6.6	**	9.1	1.7	**
	CON	3.9	−12.4	**	0.8	2.2	NS	5.1	−9.9	**	12.4	−5.9	**	6.4	−1.2	**
Nuclear/Coal	PRO	−6.8	−10.5	**	7.1	1.8	**	−2.7	3.2	**	4.4	−6.9	**	9.1	−5.0	**
	CON	3.9	−10.7	**	0.8	−1.6	NS	5.1	3.4	NS	12.4	−5.6	**	6.4	−5.8	**
Nuclear/Oil	PRO	−6.8	−8.8	*	7.1	5.5	NS	−2.7	4.7	**	4.4	−3.5	**	9.1	1.3	**
	CON	3.9	−6.6	**	0.8	4.0	*	5.1	9.1	**	12.4	−0.9	**	6.4	−0.8	**
Solar/Hydro	PRO	−10.7	−12.2	NS	3.9	6.1	NS	−11.7	−10.1	NS	−7.6	−6.6	NS	5.9	1.7	**
	CON	−10.5	−12.4	*	2.6	2.2	NS	−12.6	−9.9	*	−9.5	−5.9	**	6.5	−1.2	**
Solar/Coal	PRO	−10.7	−10.5	NS	3.9	1.8	NS	−11.7	3.2	**	−7.6	−6.9	NS	5.9	−5.0	**
	CON	−10.5	−10.7	NS	2.6	−1.6	**	−12.6	3.4	**	−9.5	−5.6	NS	6.5	−5.8	**
Solar/Oil	PRO	−10.7	−8.8	*	3.9	5.5	NS	−11.7	4.7	**	−7.6	−3.5	**	5.9	1.3	**
	CON	−10.5	−6.6	**	2.6	4.0	NS	−12.6	9.1	**	−9.5	−0.9	**	6.5	−0.8	**
Hydro/Coal	PRO	−12.2	−10.5	*	6.1	1.8	**	−10.1	3.2	**	−6.6	−6.9	NS	1.7	−5.0	**
	CON	−12.4	−10.7	*	2.2	−1.6	**	−9.9	3.4	**	−5.9	−5.6	NS	−1.2	−5.8	**
Hydro/Oil	PRO	−12.2	−8.8	**	6.1	5.5	NS	−10.1	4.7	**	−6.6	−3.5	**	1.7	1.3	NS
	CON	−12.4	−6.6	**	2.2	4.0	NS	−9.9	9.1	**	−5.9	−0.9	**	−1.2	−0.8	NS
Coal/Oil	PRO	−10.5	−8.8	*	1.8	5.5	**	3.2	4.7	NS	−6.9	−3.5	**	−5.0	1.3	**
	CON	−10.7	−6.6	**	−1.6	4.0	**	3.4	9.1	**	−5.6	−0.9	**	−5.8	−0.8	**

*Difference in mean values significant, $p < 0.05$.
**Difference in mean values significant, $p < 0.01$.
NS, difference in mean values not significant.

than solar energy but about the same as hydro. The CON group did not distinguish amongst these three energy sources with respect to economic benefits, although they did believe that both solar and nuclear energy were significantly more likely to lead to technological developments than was hydro.

Hydro As Compared to Solar Energy

The PRO nuclear group only distinguished between hydro and solar energy with respect to the question of future technological developments, solar energy being rated significantly more positive. The CON group viewed these two energy sources as being significantly different on all but the economic benefits dimension. That is, the CON group believed that solar energy was less likely to lead to environmental risk and psychological/physical risk but more likely to lead to indirect risks and technological developments.

Coal As Compared to Oil

Both groups believed that oil was more likely to lead to economic benefits and future technological developments than was coal, and that oil was also more of an indirect risk and psychological/physical risk. However, while those who were PRO nuclear energy believed that coal and oil posed equal environmental threats, those in the CON group believed oil to be significantly worse in this respect than coal.

5 CONCLUSIONS

This report has described an analysis of the Austrian public's beliefs about five energy options, and their overall attitude to each energy source. Attitudes were shown to be polarized only in the case of nuclear energy; and, regardless of their position on nuclear energy, the members of the public who participated in the survey were most favorable toward the renewable sources hydro and solar energy. The public sample *as a whole* was least favorable to nuclear energy. Those who were PRO nuclear energy, like the rest of the sample, were most favorable toward hydro and solar energy, but they were least favorable toward the fossil fuels; their attitudes toward nuclear energy were thus intermediate (on average) between their views on the renewable and the fossil sources. Given this widespread preference for hydro and solar energy it is worth emphasizing that in Austria, as elsewhere, suitable large-scale solar systems are not commercially available. Further, the attitudes toward hydro-power probably reflect favorable experience with this source, whose potential in Austria has already been developed to an extent where additional projects could not make a significant contribution to national electricity needs. Of the options studied here, only coal, oil, and nuclear energy are viable possibilities for appreciable near-term increases in Austrian electricity-generation capacity.

Austria's first nuclear power plant, a 730-MWe facility at Zwentendorf near Vienna, has been completed; however, due to adverse public reaction, and as a result of a referendum (November 1978) in which the Austrian electorate decided against the use of nuclear energy, this plant will not become operational. During the construction of the Zwentendorf plant the Austrian government sponsored a public information campaign (in late 1976 and early 1977) intended to open up debate on energy options to the general public, and the publicity given to articulate pressure groups dramatically polarized opinions with respect to the intended nuclear energy program; the resulting controversy led directly to the public referendum (Hirsch 1977).

Although the findings described here are for only a small sample of the

Austrian public, the in-depth analysis of beliefs about the different energy options can make some contribution to understanding the Austrian dilemma. This report focuses on beliefs which are relevant to a *comparison* of energy systems, but, in view of the existing controversy, also explores the perceptions of those individuals shown to be PRO or CON nuclear energy in an attempt to define the crucial differences.

NOTES

1. A related study of the beliefs underlying voting behavior in a nuclear energy referendum in the USA has also been published in this series (Bowman *et al.*, 1978).
2. The particular attitude model used in this series of reports is that developed by Fishbein and his co-workers (see Fishbein and Ajzen 1975). The way in which evaluations and belief strengths are combined to estimate attitude can be stated formally:

$$A_o \approx \sum_i^n b_i e_i$$

where

A_o = the attitude toward the object o

b_i = the strength of the belief which links the attitude object to attribute i

e_i = the evaluation of attribute i

n = the number of salient beliefs, i.e., those currently within the span of attention

3. The method used was principle components analysis of the correlation matrix followed by Varimax rotation. This technique produces underlying dimensions which do not correlate with each other (orthogonal factors).
4. The three-mode factor analysis was based on a three-way decomposition of the raw crossproducts matrix, followed by DAPPFR rotation (Direct Artificial Personal Probability Factor Rotation; R.L. Tucker, Personal Communication 1978), a method which produces oblique (correlated) factors; the intercorrelations between the factors were, however, low.
5. The 50 individuals with highest scores on the semantic differential measure of attitude toward nuclear energy.
6. This ANOVA was 2 × 5 × 5: group membership (PRO/CON) × belief dimension (5 belief dimensions derived from the factor analysis) × energy sources (nuclear energy, coal, oil, hydro-power, solar energy).

REFERENCES

Bowman, C.H., M. Fishbein, H.J. Otway, and K. Thomas (1978) The Prediction of Voting Behaviour in a Nuclear Energy Referendum. RM-78-8. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Fishbein, M., and I. Ajzen (1975) Belief, Attitude, Intention, and Behaviour: An Introduction to Theory and Research. Reading, Massachusetts: Addison-Wesley.

Hirsch, H. (1977) The "Information Campaign on Nuclear Energy" of the Austrian Government. Paper presented at the International Conference on Nuclear Power and its Fuel Cycle. Salzburg, Austria.

Osgood, C.E., G.J. Suci, and P.H. Tannenbaum (1957) The Measurement of Meaning. Urbana, Illinois: University of Illinois Press.

Otway, H.J., and M. Fishbein (1976) The Determinants of Attitude Formation: An Application to Nuclear Power. RM-76-80. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Otway, H.J., and M. Fishbein (1977) Public Attitudes and Decision Making. RM-77-54. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Otway, H.J., D. Maurer, and K. Thomas (1978) Nuclear Power: The Question of Public Acceptance. Futures 10:109–118.

Tucker, R.L. (1966) Some Mathematical Notes on Three-Mode Factor Analysis. Psychometrika 31:279–311.

ECONOMIC–DEMOGRAPHIC SIMULATION MODELS: A REVIEW OF THEIR USEFULNESS FOR POLICY ANALYSIS

Warren C. Sanderson

SUMMARY

This paper assesses the usefulness of economic–demographic simulation models for policy analysis, emphasizing in particular the relevance of the current state of the art for agricultural development planners. A critical review of eight models defines the range of questions that can be answered with particular models, evaluating the reasonableness of their specifications and the probable quality of their performance. Suggestions concerning further research are also provided.

The primary function of economic–demographic simulation models is to ascertain the quantitative importance of indirect effects of changes in the economic or demographic environment. For example, governmental policies concerning credit availability, which have a direct effect on the rate of growth of agricultural productivity, will have an indirect effect on rural population growth and rural to urban migration. A clarification of such interactions between demographic and economic phenomena is an essential ingredient of an enlightened development planning process.

The five "second-generation" economic–demographic simulation models reviewed in this paper are the FAO model, the Bachue–Philippines model, the Simon model, the Tempo II model, and the Kelley, Williamson, and Cheetham model. The main conclusion of the review is that although none of these models in their present form can offer reliable advice to agricultural policy makers, they may be useful as aids in teaching government officials about the potential long-run consequences of their decisions. Two third-generation models, the Adelman–Robinson model and the Kelley–Williamson representative developing country (RDC) model are also reviewed. Neither of these two models has a significant demographic component, but they are of interest because future economic–demographic simulation models are likely to be constructed around their fundamental concepts.

1 INTRODUCTION

This is a report on the current state of the art in modeling economic–demographic interactions, with added emphasis on the implications of this work for agricultural development. The god of manuscripts of this sort is undoubtedly Janus, one of whose faces is directed at past research, while the other points the way to future studies. In the spirit of Janus, this paper has two aspects – first, a critical review of selected economic–demographic models of development and second, a set of suggestions concerning further research.

Over the past decade and a half, the population of economic–demographic simulation models of the process of development has virtually exploded. The first such model appeared in 1963, and even by 1970 their number could be counted on one hand. Currently, however, although a complete count is difficult to make, there must be several dozen of these models in existence. Thus, policy makers who currently do not have economic–demographic planning models at their disposal will increasingly want to know whether there are any models that are suitable for their purposes, and those who do have such models at hand will increasingly want to know how their model compares with other similar planning tools. It is to these people that this paper is addressed.

Before we begin the review of the models, however, a brief discussion of their nature and purpose is in order. The primary function of economic–demographic simulation models is in ascertaining the quantitative importance of the *indirect* effects of changes in the economic or demographic environment. The models are not designed to give detailed guidance to policy makers about the direct effects of their decisions. For example, an official interested in increasing agricultural productivity will not find any of the models reviewed here very helpful. Expert advice from individuals specializing in agricultural policies, agronomy, animal husbandry, and pest control is likely to be of far greater use to him. Similarly, a policy maker who is interested in reducing rural fertility will not get much detailed guidance on how to do so from any of the models. For that purpose, he would be better served by consulting public health personnel. The models in this paper are not constructed to address such questions. Their usefulness is strictly limited to a different set of concerns – *interactions* between diverse phenomena.

Policy makers who are concerned, for example, with increasing agricultural productivity may well be interested not only in the direct effects of certain policies on agricultural output, but also in the indirect effects of those policies on rural population growth and rural–urban migration. Policy makers interested in demographic issues, such as policies concerning expenditures on family planning or policies affecting internal migration, may well be interested in the indirect effects of these policies on economic development. It is in such connections that the models may be legitimately employed because they can alert planners to indirect effects that can significantly reduce or enhance the thrust of their policies. The usefulness of these models does not arise from any of their aspects taken in isolation, but rather from the *interactions* between their various components. The proper role of economic–demographic planning models, then, is a modest one. Such models provide the policy maker with one tool, among the many he needs, to make sound judgments about the alternatives available to him.

Viewed in this light, questions concerning what is included in and what is excluded from economic–demographic simulation models can be answered with greater clarity. These models need to be sufficiently articulated to address major policy issues. They need to be strong in the area of economic–demographic interactions, but can be sketchy in certain details relating to the economy and the demography of the country.

Granted that economic–demographic simulation models have a modest place among the tools of development planning, the question naturally arises as to how well existing models perform the limited role for which they are useful. Unfortunately, this straightforward and important question has no simple answer. The models reviewed here are designed to understand the long-run pace and character of the development process, not short-term economic or demographic changes. To test directly whether the quantitative implications of a given model were correct in even one instance would require a lengthy experiment and a substantial amount of analysis of the resulting data. It is possible conceptually to test the models over some past era, but as a practical matter this is generally impossible because historical data are not available and in many cases the relevance of the model specifications for historical anlysis is dubious. Therefore, in evaluating economic–demographic simulation models the direct approach of testing their implications against reality is not feasible.

There is, however, the possibility of indirectly reviewing the usefulness of existing models. To understand how this can be done requires a brief discussion of the nature of those structures. Each of the models is composed of three related parts:

1. A list of parameters and exogenous and endogenous variables
2. A list of equations relating the exogenous and endogenous variables and the parameters
3. A set of values for the exogenous variables and parameters as well as for the initial values of the endogenous variables

The first component defines the set of questions that can be answered by using a particular model. The changes in any set of endogenous variables due to alterations in any exogenous variables and parameters may properly be studied. Since the models have different focuses, it is natural that their lists of exogenous and endogenous variables should differ. Unfortunately, the lists of exogenous and endogenous variables and parameters are bound to be a disappointment to those interested in agricultural planning. The models, with the exception of the one developed by the UN Food and Agriculture Organization (FAO), cannot address many of the questions of great importance for policy purposes.

The third component, the actual figures that are utilized in the versions of the models reviewed here, is not discussed in this paper. There are two reasons for this. First, these data are almost uniformly of poor quality. Indeed, many of the numbers used in the simulations are nothing more than educated guesses. Although guesstimates and approximations are often sufficiently accurate for the purposes of simulation, there is no easily available method for ascertaining whether one set of poor data is preferable to another set of poor data. The second reason for not discussing the input data here is that policy makers who are potentially interested in using a given framework are not as concerned about the figures in any given application as they are about whether the structure of the model can profitably be applied in their particular case.

The second component, the equations, forms the heart of any economic–demographic simulation model. Evaluating the equations provides an indirect basis for judging the likely performance of models. The specifications of the equations can be rated according to three criteria:

1. Do they allow the questions posed by the model to be answered in a meaningful manner?
2. Are they plausible?
3. Are they technically correct?

The first criterion is the most subtle of the three. Suppose for a moment that one important question to be answered by a particular model concerns the relationship between the rate of population growth and the rate of *per capita* income growth. Further, let the production function that relates aggregate output Y to the factors of production land A, labor L, and capital K have constant returns to scale. We may write

$$Y = f(L, K, A \mid T) \qquad (1.1)$$

where T represents the technology at any moment in time. Now, if the model assumes that the rates of growth of the capital stock and the stock of land are independent of the rate of growth employment, that the rate of technological change is also independent of the rate of growth of employment (although it may depend on the rate of growth of the capital stock or the stock of land), and that the labor force/population ratio is constant, then decreasing the rate of growth of the population always increases the rate of growth of income per capita.[1] This conclusion obtains regardless of the parameter values. Indeed, it even holds for any constant returns to scale production function. If one did not know that this conclusion was built into the basic structure of the model, one might even be tempted to demonstrate how "robust" it was to parameter changes.

Such a model would not allow the question of the relationship between population growth and per capita income growth to be answered in a meaningful way because the direction of that association is assumed in the specification. Although the frameworks reviewed here are considerably more complex than the simple example above, some of them come quite close to postulating the results of their analyses. A number of such cases are discussed below.

The second principle on which to judge a specification is its plausibility. For example, one of the models assumes that agricultural output depends solely on the number of people employed in the agricultural sector and is independent of the agricultural capital stock and such material inputs as fertilizer, seeds, and water, while another model assumes precisely the reverse. It is implausible, however, to assume that either the marginal product of agricultural labor or agricultural capital is zero in the long run even if one or the other were true in the short run. The results of a model that contains implausible specifications of important relationships should be treated with caution by policy makers. Many, but not all, of these implausible specifications are described in detail below.

The third principle on which to evaluate one or a set of equations is their technical correctness. For example, in one of the models reviewed, two sets of demographic variables related to marriage and fertility are inconsistent with each other. The same model determines the output prices used in its consumption equations inappropriately. Such technical errors should be corrected before its simulation results are seriously considered by policy makers. Several such technical mistakes are revealed in the model reviews below.

Although the implications of the economic-demographic simulation models cannot be directly tested, a good idea of their likely performance can be gathered from an evaluation of their structures. Chapter 2 provides a summary of such evaluations for the seven models reviewed here.

2 OVERVIEW

This paper reviews five second-generation economic–demographic simulation models[2] and assesses their usefulness for agricultural policy formation in developing countries. The main conclusion of the review is that none of these five models in their present form can give serious guidance to an agricultural policy maker. Two third-generation simulation models, those of Adelman and Robinson (1978) and Kelley and Williamson (1979), are also reviewed here. Neither of these two models has a significant demographic component. They are interesting from our present perspective because future economic–demographic simulation models are likely to be constructed using their frameworks. Policy makers interested in economic–demographic simulation models would be well advised to begin with the Kelley–Williamson (1979) model and to expand it where necessary to address issues of relevance to their country.

2.1 THE FAO MODEL

The Food and Agriculture Organization model of Pakistan is the only model reviewed here that has any relevance to agricultural policy questions. The model consists of four segments: agricultural output, nonagricultural output, employment, and demography. Each of these segments and the model as a whole are constructed very simply. Indeed, in concept, the FAO model is the simplest of all the models reviewed. This simplicity is both its principal advantage and its main disadvantage. It allows, on one hand, a complete model to be built with very little actual data. This is a necessary characteristic of any model that is designed for widespread use in less developed countries. On the other hand, however, the simplicity weakens the credibility of the model's implications.

Four types of agriculture are distinguished in Pakistan: small-scale farming in rainfed regions, large-scale farming in rainfed regions, small-scale farming in irrigated regions, and large-scale farming in irrigated regions. In irrigated farming regions a certain amount of acreage is assumed to be withdrawn from cultivation each year and a policy-determined amount of land reclaimed. The government can, at a fixed cost per acre, redistribute land to small farmers or consolidate it into larger farms. In addition, government policy determines the amounts of investment and intermediate inputs such as fertilizer going to agriculture. The specification of the agricultural production process, however, is so simple that the results may not be meaningful. For example, since the production process assumes a constant marginal product of capital (i.e., agricultural capital never encounters diminishing returns even with a fixed quantity of land), it is likely that the optimum agricultural strategy for the government is to concentrate all agricultural investment in one of the four types of farming.

There are a number of omissions from the agricultural submodel that limit its usefulness. Foremost among these is the almost complete lack of attention to technological progress and its differential effects on various forms of farming. Another important omission is any consideration of the agricultural labor force. While it may be argued that labor is a redundant resource in agricultural Pakistan today, it hardly seems useful to assume that no development policy over the course of two or three decades will result in agricultural labor having a positive marginal product.

Output in the nonagricultural sectors[3] is similarly treated with extreme simplicity. Government policy is assumed to determine investment allocations in the modern sector, and all production processes are assumed to be characterized by constant marginal products of capital. Given the fixed relative prices implicit in the FAO model, nonagricultural output is maximized when the government invests in only that sector with the highest marginal product. Again, the quantities of labor used in the nonagricultural sectors of the economy have no influence on their levels of output. Further, the model has no demand functions for the various nonagricultural products except construction. Technological change embodied in new capital is allowed in the nonagricultural sector, but is not implemented in the Pakistani simulations. Disembodied technological change is not allowed to occur.

Besides migration and the specification that the country has a fixed budget in each year to spend on investment, the agricultural and the nonagricultural sectors are essentially unconnected in the FAO model. Migration is taken as depending on, among other things, the relative output–labor ratios in the agricultural and nonagricultural areas. This is taken as a proxy for the relative nonagricultural and agricultural wage rates, which are not determined. How good a proxy it is remains an open question.

The demographic submodel is not implemented in the Pakistani case. Instead, various assumptions are made about population growth rates. The educational system is also omitted from the present model, which may be just as well, since education is assumed to affect only fertility.

In short, the FAO model in its present form is simple enough to implement but not yet complex enough to be realistic. This is a common difficulty with these models, but the FAO model is the most simplistic of the models reviewed here. In particular, the specification of the agricultural sector is simplified to the point of unreality. Policy makers should, therefore, be wary of using the FAO model to guide the formulation of agricultural policy even though it is one of the few models that deals, even in modest detail, with agriculture.

2.2 THE BACHUE–PHILIPPINES MODEL

The Bachue–Philippines model differentiates 13 sectors, among which are domestic food crops, export crops, livestock and fishing, and forestry. This makes Bachue by far the most disaggregated second-generation model and the one with the most specificity in regard to agricultural outputs. Bachue is unlike the other models in that in most of its simulation runs the rate of growth of aggregate output is assumed to be exogenous. The model, therefore, is not designed to answer questions concerning the effects of policy decisions on the rate of economic growth. The focus of the model instead is on the distribution of income. Thus, Bachue is most useful in analyzing the effects of changes in the economic and demographic environment on the distribution of income in those cases where the changes themselves and the resulting alterations in the income distribution have little or no effect on the rate of economic growth. Another respect in which Bachue is unique among the models reviewed here is in its specification of the relationships between inputs and outputs. Except in the case of domestic food production, neither capital nor labor inputs play any role in the derivation of sectoral output levels. The quantity of domestic production in each sector is determined essentially by demand conditions. The quantities of the factors of production are calculated only after output levels are known.

The heart of the economic segment of the Bachue model is a 13 × 13 input–output matrix for 1965 that is assumed to remain unchanged over the simulation period. In order to avoid simultaneity, the final demand for the output of each of the 13 sectors is assumed to be predetermined in each year. Given the vector of final demands, the input–output matrix is used to compute the quantities of output produced by each sector. The usual procedure, given an input–output matrix and a vector of final demands, is to subtract competing imports from the vector of final demands in order to determine the vector of gross outputs. Instead of using this procedure, the model contains a system of simultaneous equations that jointly determine imports and gross outputs. This is a good idea, but the specific equations yield the implication that whenever a sector's exports increase (say, because of an increase in productive capacity), the sector's imports also increase. This hardly seems like a plausible assumption to make concerning all sectors of the economy.

Value-added per unit of output in current prices in the thirteen sectors are allowed to take on only two values, one for goods dominantly produced in rural areas and one for goods mainly produced in urban areas. The ratio of the two value-addeds is determined by the relative supply of and demand for domestically produced foods. On the supply side, labor productivity growth in the domestic food crop sector depends mainly on an exogenous (policy) parameter and to a limited extent on the rural–urban value-added ratio. Labor productivity and employment in the domestic food crop sector alone – capital and intermediate inputs play no explicit role here – determine the supply of domestic foodstuffs. The demand for domestic foods is computed as described above. When demand and supply are not identical in a given year, the relative value-added in the model changes in the following year. Current imbalances are eliminated through foreign trade. The two value-addeds for 1965, however, are inconsistent with those used to create the 1965 input–output table. Further, the output prices derived from the value-addeds are not appropriately used in the deflation of quantities of output measured in monetary units. A detailed procedure for correcting these problems is contained in section 3.2. The level of investment, like the level of aggregate output, is treated as exogenous in most of the simulations of the Bachue model. This has certain immediate implications: saving is essentially unrelated to investment, and investment is unrelated to both the level and growth rate of output in the Bachue model. Further, a technical problem also arises because of the exogenous nature of investment – how to allocate investment funds to sectors whose growth rates have already been determined by the input–output analysis. In Bachue, this is accomplished by using a set of fixed incremental capital–value-added ratios. Unfortunately, nothing guarantees that the aggregate amount of investment so computed equals the exogenous level of investment. This inconsistency is reconciled by an *ad hoc* adjustment of investment demands.

The income distributions in the model are based on (a) the distributions of employment not only across sectors but also with regard to self-employment and wage employment in most of the sectors and (b) the average annual incomes of the people in each category of employment. The methods of deriving the requisite numbers here are complex and in many instances not totally convincing. For example, the average annual incomes are incorrectly computed because of an error in moving from value-added in constant prices to value-added in current prices. To obtain distributions of household income from data on the distributions of employment and average annual incomes requires the transformation of information on the incomes of individuals to information on the incomes of households. Whether the complex procedure used to do this would yield reasonable approximations to true income distributions given correctly computed input data is difficult to ascertain.

The demographic portion of the Bachue model is both reasonably simple and sophisticated. Age-specific mortality and marital fertility rates are computed, as well as age-specific proportions of women currently married and age-specific numbers of people enrolled in school. There are, however, technical errors in this segment of the model as well. For example, the age-specific marital fertility rates and proportions of women currently married are inconsistent with the gross reproduction rate also derived in the model. Once the technical errors discussed in section 3.4 are corrected, the demographic segment of Bachue would easily be superior to those in the other models reviewed here.

The Bachue model has both strengths and weakness. Its attention to the details of the distribution of income and demographic processes is surely to be applauded. On the other hand, the economic portion of the Bachue model is, currently, quite weak, particularly with regard to the relationship between income distribution and economic development. Even some of the details of the income distribution process are technically incorrect. The model will be considerably strengthened when the technical errors are corrected and when serious attention is paid to making output growth and investment endogenous.

2.3 THE SIMON MODEL

The model of economic-demographic interactions created by Simon differs considerably from the other models reviewed here. Like the Kelley, Williamson, and Cheetham and Tempo II models, it has an industrial sector and an agricultural sector. Unlike those models, it was not developed to be applied in particular contexts, but rather as a tool for the study of the effects of population growth on economic development. This focus leads the Simon model to concentrate on relationships that run from population growth to economic development rather than from economic development to population growth. Perhaps the most unusual feature of the Simon model, though, is that in each year total output and total hours of work are chosen so as to maximize the country's social welfare function. This is one approach to making the hours of work done by the inhabitants of a given country endogenous. A more conventional and probably preferable approach to the same end would have been to specify labor supply functions separately in each of the two sectors of the economy. The social welfare function in the Simon model is not a stable one, but rather one that shifts around with changes in per capita income and the dependency rate. Whether a country can realistically be modeled as maximizing a social welfare function and whether that function can reasonably be characterized as shifting in the manner assumed by Simon are at best open questions and at worst unanswerable ones. A policy maker who does not know his country's social welfare function should not think seriously of using the Simon model.

The industrial and agricultural sector are both characterized by Cobb-Douglas production functions that allow for neutral technological progress. Output in each sector is produced using three factors of production: labor in the sector, capital (including land) in the sector, and the country's entire stock of social overhead capital. The elasticity of output with respect to social overhead capital in the two production functions is unity. Social overhead capital is assumed to grow at some fixed fraction of the rate of growth of the labor force! Thus, Simon sees more rapid population and hence labor force growth as increasing the rate of output growth, in part, by its effect of increasing the rate of accumulation of social overhead capital.

The agricultural capital stock in the Simon model is augmented annually by a quantity of investment that depends on the agricultural labor–capital ratio and the stock of social overhead capital in the previous period. The industrial investment specification, on the other hand, is apparently in error because it implies that net investment in industry is always negative. Technological change in the agricultural sector is assumed to proceed at a steady one-half of one percent per year. Technological change in the industrial sector is assumed to occur at a somewhat slower pace. Precisely how much more slowly depends upon the rate of growth of industrial output. For example, if industrial output is growing at ten percent per year, then technological progress occurs at a rate of three-tenths of one percent per annum; if it is growing at one percent per year, then technological progress occurs at a rate of one-tenth of one percent per annum. The rationale for the assumption of slower technological progress in industry than in agriculture is not stated in the Simon model.

The distribution of output between the two sectors of the economy in period t (assuming invariant relative prices) is assumed to depend upon the level of per capita income in period $t-1$. As per capita income increases, it is assumed that the country automatically becomes more industrialized. There are no demand equations in the Simon model, no specification of the savings rate, no migration rate formulation, no educational structure, nor any information about the distribution of income between labor and capital.

The Simon model is an attempt at obtaining a simulation model that can be used to ascertain the effects of population growth on economic development. Unfortunately, the model makes a number of unconvincing structural assumptions and may contain outright economic errors. No policy maker should be influenced by the Simon model in its present form. Nor is this model a useful framework to develop for policy purposes. There are no interesting agricultural policy questions that can be addressed in the context of the present Simon model.

2.4 THE TEMPO II MODEL

Tempo II is a two-sector model that distinguishes a rural subsistence sector from an urban industrial sector. Industrial output is assumed to be generated by a Cobb–Douglas production process that allows for neutral technological change to occur at a constant rate over time. The inputs are assumed to be unskilled labor, skilled labor, and capital. Of all the models considered here, only Bachue and Tempo II allow education to enhance the productivity of workers.

The output of the agricultural sector, however, is assumed to be produced by labor alone, and no technological change is allowed to occur in agriculture over a simulation period of twenty to thirty years. Thus, agricultural land and capital play no role in the development process. Further, there is no social overhead capital either in the rural area or in the urban area. It is clear, then, that in the world of Tempo II, policy makers cannot increase agricultural output by teaching farmers to employ new techniques, by educating farmers generally, by increasing the capital intensity of agriculture, or by building rural social overhead capital. Indeed, there are no policies of agricultural development that are enlightened by Tempo II.

The outputs of both sectors in period t depend upon the quantities of inputs used in production in period $t - 1$. This rather odd specification ensures that the physical outputs in any period are essentially predetermined. Relative output prices are held fixed at unity over the simulation period – a weak assumption made in all the second-generation models except Bachue and the Kelley, Williamson, and Cheetham model – and income in any period is set equal to output in that period. The government, however, is allowed to run a deficit that is covered in part by the printing of money. In that case, aggregate demand, which is simply income plus the monetized portion of the government deficit, must exceed output, causing a generalized inflation to occur. As a practical matter, all elements of aggregate demand (except expenditures on education and family planning services) are reduced proportionally until aggregate demand and supply are again in equilibrium. Tempo II is the only model reviewed here that allows a government deficit to be covered by printing money.

With disposable income held constant, private savings per capita and therefore private investment per capita in the Tempo II model are assumed to be negatively related to the size of the population. This is in direct contradiction to the specification of investment in the Simon model. Since the capital stock in the urban area is the only capital stock in the country, it is determined from a base-period capital stock estimate plus accumulated net investment.

In the agricultural sector, the entire populace is considered as working, and an infant and an adult are each counted as one unit of agricultural labor. In the urban area, the size of the skilled and unskilled labor forces are determined by applying exogenous age- and sex-specific labor force participation rates to the age- and sex-specific numbers of skilled and unskilled workers. The numbers of skilled workers employed and unemployed are assumed to be fixed proportions of the skilled labor force. The number of unskilled workers employed, however, is determined from a very dubious equation that relates this number *negatively* to the size of the unskilled labor force if the ratio of the unskilled labor force to the capital stock is fixed. In other words, if the unskilled labor force and the capital stock were both to grow at, say 2 percent per annum, unskilled employment would *decline* continuously until eventually both it and industrial output would go to zero. This is hardly a realistic specification.

Tempo II is a policy-oriented model and is especially strong in its formulation of family planning policy. It is assumed that only the government spends money on fertility control and then only in the urban area. Further, it is assumed that up to a point the cost to the government of averting a birth remains constant. After that point is reached, the cost to the government of additional births averted rises. The cost to the government of a family planning program depends on how many births the government wishes to avert. With enough money, the government can always attain its fertility control objectives. It is interesting to note in this regard that nothing but the family planning program can affect birth rates in Tempo II, and, since there can never be a family planning program in the rural area, rural fertility rates are immutable for the entire simulation period of perhaps two or three decades.

The only policy that can be sensibly studied in the context of Tempo II is the government's policy toward family planning. Unfortunately, the specification of Tempo II ensures that increases in family planning expenditures will always cause an increased per capita income whenever the cost of averting an additional birth is less than twice the per capita income of the country. Indeed in the long-run, in the Tempo II model, expenditures on fertility control could increase per capita income even if the cost of averting an additional birth were about five or six times per capita income. This result is essentially built into the Tempo II framework by assuming that population growth has no stimulating effects anywhere in the economy. If this is what a policy maker believes, then the Tempo II result on family planning follows without a simulation model. If this is not what a policy maker believes, then he would be well advised not to accept the results of the Tempo II model.

2.5 THE KELLEY, WILLIAMSON, AND CHEETHAM MODEL

The Kelley, Williamson, and Cheetham (KWC) model of dualistic economic development in Japan is by far the most economically sophisticated of the second-generation models reviewed here. It is not designed to be a policy-oriented model, but rather is a model designed to shed light on Japanese economic development. Nonetheless, the KWC model has more potential for policy analysis than any of the other second-generation models that have been reviewed. The KWC model helps one to understand the behavior of a number of interrelated time series concerning Japanese economic growth and in this sense may be considered to be the only successfully tested model reviewed here.

The KWC model divides the Japanese economy into two sectors, an agricultural sector and an industrial sector. In both sectors a CES production function is assumed, with capital and labor as the inputs. This is a more sophisticated formulation than is used in any of the other studies. The importance of this specification is twofold. First, the use of the Cobb–Douglas production functions would constrain the elasticities of substitution between labor and capital to be unity in both sectors – a highly debatable assumption. Indeed, Kelley, Williamson, and Cheetham cite evidence suggesting that the elasticity of substitution is significantly smaller in the industrial sector than in the agricultural sector. The flexibility of the CES specification is not the only reason to prefer it. Perhaps a more important reason is that it allows the incorporation of biased technological change into the model. The KWC model and the Kelley and Williamson (1979) model are the only ones reviewed here that take this vital aspect of economic development into account.

Not only does the KWC model treat the supply side of the economy sensibly, it also treats the demand side in a plausible manner. The demands for the two goods in the economy are derived from a Stone–Geary demand structure. The interaction of the demand side and the supply side of the economy, logically enough, determines the quantities of the outputs produced and their relative price. It is rather disconcerting to realize that in none of the other second-generation models reviewed were outputs determined in any meaningful way by the interaction of supply and demand, nor, with the exception of the Bachue model, were relative prices considered to be endogenous.

This last point is extremely important. Over the course of economic development the terms of trade between industry and agriculture have a tendency to change for a number of reasons. Indeed, many agricultural policies themselves could be expected to affect the relative price of agricultural output. Models that do not have endogenous relative prices are severely handicapped for policy analysis. For example, without knowing the price of agricultural output relative to the price of industrial output, it is impossible to compute the relative wages of unskilled laborers in the two sectors and, hence, essentially impossible to obtain a reasonable migration specification. Similarly, it is impossible to compute relative rates of return to capital in the two sectors. This list can be made substantially longer, but the important point to remember is that policy makers ought not to consider seriously the implications from models of economic–demographic interactions that do not contain any endogenous relative prices. Such models are likely to lead them substantially astray.

Since in the KWC model the price of agricultural goods relative to industrial goods is endogenous, it is possible to compute the incomes of laborers and the return to capital in the two sectors. It is assumed in the KWC model that all labor income is consumed and that a portion of income from capital is saved and reinvested. Two specifications of how investment is allocated between sectors are given in the KWC model. The more relevant formulation assumes that capital stocks in each sector can be derived from an estimate of the base-year stocks and cumulated net investment. Investment in a given sector depends on the sectoral distribution of savings and the relative rates of return on capital in the two sectors. If the rates of return are not too different from one another, savings are assumed to remain in their sector of origin. If the rates of return are sufficiently out of line, some savings will flow from the low-rate-of-return sector to the high-rate-of-return sector. Migration is treated similarly in the KWC model. If the wage in the industrial sector is enough greater than that in the agricultural sector to overcome the cost of migration, then people will move from rural areas to urban areas. The greater the wage gap, the greater will be the migration rate.

Although the KWC model is not policy oriented, its framework is useful for policy analysis. For example, one can test the effect of stimulating agriculture by subsidizing agricultural output or the effect of inducing greater agricultural investment by subsidizing agricultural capital formation. Further, it is straightforward in the KWC model to experiment with policies that affect the rate of bias of technological change in agriculture. The principal weakness of the KWC model in its present form is its demographic specifications. The age structure of the population, for example, is not included in the model at all, and urban and rural fertility rates are taken to be wholly exogenous. The Kelley and Williamson (1979) model, discussed below, is an extension of the KWC model. It is a useful foundation for further development, but in its present form it also lacks much demographic structure.

2.6 THE ADELMAN–ROBINSON MODEL

Two third-generation development simulation models are reviewed here, the Adelman–Robinson model of Korea and the Kelley–Williamson model of a representative developing country. These models are more sophisticated in their economic specifications than are the second-generation models. Like the Kelley, Williamson, and Cheetham model, both of the third-generation models determine output prices, factor prices, and the composition of output simultaneously.

The Adelman–Robinson model of the Korean economy differs from the other models reviewed in this paper in its time horizon. While the other models are concerned with economic–demographic interactions that occur over the course of one or more generations, the Adelman–Robinson model is concerned with a time span shorter than a decade. The focus of the Adelman–Robinson model is on questions concerning the relationships between economic growth, economic policies, and the size distribution of household income. In its concerns and in some of its details, the Adelman–Robinson model is similar to the Bachue model. It is instructive, therefore, to compare and contrast the models in order to see which specifications are most useful in various contexts.

The Adelman–Robinson model is quite large, containing over 3,000 endogenous variables. It contains equations describing the workings of Korean financial markets, both formal and informal, equations representing 29 sectors of the economy, each containing firms of 4 sizes, and equations for the functional distribution of income and for the size distribution of household income of 15 distinct groups of income recipients.

The production functions for the urban commodity-producing sectors of the economy are assumed to be Cobb–Douglas in form. Agricultural output is produced by a two-level two-input CES production function where the factors are assumed to be capital and a labor aggregate, computed using a Cobb–Douglas specification.

Most labor supplies in the model are essentially exogenous. Some endogeneity is introduced, however, for 3 of the 15 categories of income recipients.

The demand for labor is determined from a specification that assumes that all firms are profit maximizers and that, therefore, laborers are paid the value of the marginal product. Instead of computing several hundred wage rates simultaneously, the model determines only one average wage rate for each of the 15 categories of income recipients. This greatly simplifies the computational burden of such a large model. Most of the remaining wage rates in the model are assumed to be fixed multiples of one or another of the 15 wage rates. Thus, in many cases, 78 wage rates are derived from a single average wage rate.

The procedure of computing 78 wage rates as fixed multiples of a single figure computed in the model is unfortunate in the context of a model whose focus is on questions concerning the distribution of income, because it builds into the model a substantial bias in favor of the conclusion that the distribution of income is quite stable.

Survey data are used to translate the functional distribution of income produced by the economic model into the size distribution of household income. The procedure used here and in the Bachue model to perform this function are quite similar. Among the assumptions made in this portion of the model are that the income distributions in each of 15 recipient groups is lognormal and that the (log) variances of about half of these distributions are exogenous to the model. The other half of the distributions have (log) variances that are determined mainly by the fixed multipliers mentioned above. Changes in the national distribution of income in the Adelman-Robinson model, then, must come mainly from alterations in mean incomes of various groups of income recipients and from changes in the occupational composition of the labor force.

In the Adelman-Robinson model, income available for consumption is determined by subtracting from nominal income savings, taxes, and changes in the holdings of money balances. The inclusion of money balances in the model allows Adelman and Robinson to construct a formulation in which the rate of inflation is endogenous. They are certainly to be applauded for recognizing the importance of this problem for contemporary developing countries. Unfortunately, however, desired change in the stock of money holdings by various household groups is not assumed to be a function of changes in that group's economic conditions, but rather to be an exogenous proportion of the aggregate change in the money stock.

Given income available for consumption, the commodity composition of consumption expenditure is based on a system of demand equations in which income and price elasticities are assumed to be constant during any given period. These elasticities are adjusted from period to period for the sake of accounting consistency.

Migration from rural to urban areas is treated very simply in the Adelman-Robinson model. The rate of migration is assumed to depend on the difference between the average incomes of workers in the sectors that are assumed to send the migrants and the average incomes of workers in the sectors that are assumed to receive the migrants. There is no mention in the model of any consideration of cost-of-living differences between urban and rural areas, nor do the characteristics of the income distributions in the urban and rural areas play any role in the migration decision.

The financial sector of the economy is specified in more detail in the Adelman–Robinson model than in any of the other models reviewed here. The function of the financial sector in the model is to allocate investment funds to the various sectors of the economy based on expectations of their future sales, output prices, factor prices, and profitability. The formulation in the model is a detailed one, which takes account of both the formal financial sector and the "curb" market.

The Adelman–Robinson model is a pioneering piece of research that will undoubtedly have a substantial influence on future model builders. In particular, the concern of Adelman and Robinson with the size distribution of household income in addition to the functional distribution of income has already influenced the character of the Bachue model and will certainly influence the shape of many future models as well. It is somewhat unfortunate in this connection that some of the specifications concerning the distribution of income in the Adelman–Robinson model are weak. I am confident, however, that further work in the area will strengthen them.

2.7 THE KELLEY–WILLIAMSON REPRESENTATIVE DEVELOPING COUNTRY (RDC) MODEL

The Kelley–Williamson representative developing country model is an extension of the KWC model discussed above. In the RDC model, as in the KWC model, output prices, factor prices, and the composition of output are all endogenous and simultaneously determined. There are eight sectors in the RDC model in contrast to the two sectors in the KWC model. The chief difference between the models, however, is not in the number of sectors but in the characteristics of the sectors. The RDC model distinguishes between manufacturing, agriculture, urban modern services, urban traditional services, rural traditional services, urban high-cost housing, urban low-cost housing, and rural low-cost housing. The first two of these outputs are assumed to be tradable both internationally and between urban and rural areas, and the third is assumed to be internally tradable, but not internationally tradable. In the remaining five sectors, however, outputs are assumed to be consumed only in the area in which they are produced. Thus, the outputs of a majority of sectors in the RDC model are neither internationally or interregionally tradable. The inclusion of internally nontradable goods differentiates the RDC model from all the other models reviewed here and permits the RDC model to capture aspects of the development process that are more difficult or impossible to study in the other models.

The production functions used to represent the two urban modern sectors (manufacturing and modern services) are two-level CES functions. These functions are consistent with a body of development literature that stresses that skilled labor and physical capital are complementary inputs. The demand for intermediate inputs purchased domestically is assumed to be derived from a set of fixed coefficients, as is the demand for intermediate inputs purchased from abroad. While the two-level CES production functions allow for factor-augmenting technological progress, for unbalanced technological change across sectors, and for complementarity as well as substitutability between the factors of production, the fixed coefficients allow neither for any intermediate input-saving technological change nor for any substitutability of any sort. The fixed-coefficient assumptions could introduce a substantial bias into the output of long-period simulation runs.

The production function representing agriculture is Cobb–Douglas in form with added fixed-coefficient assumptions concerning intermediate inputs. The outputs of the traditional service sectors are assumed to depend only on their levels of labor inputs, and the outputs of the housing sectors are assumed to depend only on the stocks of the various sorts of housing.

Given that capital stocks and aggregate labor supplies are predetermined in any given year and that all factors of production are paid the value of their marginal product, wage rates and the structure of employment are determined conditional on the following three assumptions: (a) unskilled labor in the rural sectors is perfectly mobile between those sectors; (b) skilled labor in the urban modern sectors is perfectly mobile between those sectors; and (c) unskilled labor in the urban areas is perfectly mobile between the two modern sectors and always is paid a constant percentage more than unskilled labor in the urban traditional service sector.

The RDC model makes an important advance over the other models discussed here in its formulation of the structure of savings and consumption. For this purpose, the model utilizes the newly developed extended linear expenditure system (ELES). The advantage of this specification – and it is indeed a substantial one – is that savings and consumption decisions are made in a unified framework and influenced in a consistent manner by income and relative prices. For example, the ELES system framework savings rates may be affected by alterations in the price of food. No other model considered here can capture such effects.

The allocation of investment funds in the RDC model is performed by a dual financial structure. Finance for investment in housing is assumed to originate only in the sector in which the housing is demanded. Further, housing finance is the first-priority use for savings. Only if there are funds left over after housing needs are met is there any nonhousing investment. The financial market in which nonhousing investment funds are allocated is assumed to be reasonably efficient, so that differences in marginal rates of return between sectors are minimized.

There are two aspects of the dynamic portion of the model that deserve mention here: migration and the rate of growth of the skilled labor force. The migration formulation in the RDC model is quite strong. Migrants are motivated to move from rural areas to urban areas because of real income differences. In computing these differences the rural migrants are assumed to take into account both differences in the cost of living between the parts of the country and the income distribution in the urban area and the associated probabilities that they would be able to obtain specified income levels.

Migration, then, plays a far more important role in the RDC model than it does in the other models. Migration in the RDC model affects the level of nonhousing capital formation by affecting the demand for housing and housing finance. On the other hand, migration also causes a set of changes in relative costs of living, which, in turn, reduces migration. No other model has been able to capture the interactions of forces such as these.

In most of the models reviewed here, the rate of growth of the skilled labor force was taken either to be completely exogenous or to depend on governmental policy with respect to expenditures on education. The RDC model, however, takes a position, first used, to my knowledge, by Edmonston *et al.* (1976), that there is an additional source of skilled laborers. When it becomes profitable for them to do so, firms can also train skilled workers. This is, I believe, an important feature to build into any long-run economic–demographic simulation model.

The chief disadvantage of the RDC model from the point of view of a policy maker interested in economic–demographic interactions is that the model in its current state is demographically underdeveloped. The authors discuss some possible demographic extentions of their model, and these would certainly be useful.

Policy makers interested in the construction of an economic–demographic simulation model for their own country would be well advised to begin with the framework of the RDC model and to add to it enough relevant detail to enable it to address questions of interest to them. For example, a policy maker may wish to add some material on income distributions from the Adelman–Robinson model, material on family planning and education from the Tempo II model, and some material on marriage rates from the Bachue model. It is crucial, however, that the additions be made on a consistent and realistic foundation – and this is exactly what the RDC model is.

3 THE BACHUE- PHILIPPINES MODEL

The Bachue-Philippines model, constructed with support from the International Labour Organization, is the largest and most ambitious of the second-generation models. It is composed of roughly 250 behavioral equations and identities (some in matrix form) and contains over 1,000 economic variables and over 750 demographic variables. One might expect a model of this size also to be one of unusual sophistication throughout, but this is not the case with the Bachue model. Instead, it is focused on issues relating to the distribution of income and employment. This is not to say that other matters have been completely ignored. Far from it: the model deals with a wide variety of additional issues. The treatment of those issues, however, is often extremely simplified, in contrast to the detailed consideration given to questions concerning the distribution of earnings and employment. Even in a model as large as Bachue, hard decisions have to be made concerning which aspects of reality should be emphasized and which should not.

3.1 DETERMINATION OF THE LEVELS OF GROSS AND NET OUTPUTS

The heart of the process of output determination in Bachue is a 13-sector input-output table based on 1965 data. The sectors are domestic food crops, export crops, livestock and fishing, forestry, mining, modern consumer goods, traditional consumer goods, other manufacturing, construction, transportation and public utilities, modern services and wholesale trade, traditional services and retail trade, and government services. In any year, say year t, the corresponding vector of final demands for these 13 sectors, $F(t)$, is assumed in the Bachue model to be predetermined. In other words, consumption, investment, and government expenditures in year t are assumed to be independent of output levels and income in year t. This is an important assumption in the model, and we shall return to it several times in the discussion below. The usual procedure, given an input-output matrix and a vector of final demands, is to subtract competing imports from the vector of final demands and to premultiply the difference by the inverse of the Leontief matrix to obtain the corresponding 13 x 1 vector of gross outputs. This procedure is shown in equation (3.1):

$$X^*(t) = (I-A)^{-1}[F(t)-Im(t)] \tag{3.1}$$

where $X^*(t)$ is the 13 x 1 vector of gross outputs in year t, I is a 13 x 13 identity matrix, A is the 13 x 13 input–output matrix, $F(t)$ is the 13 x 1 vector of final demands in year t, and $Im(t)$ is the vector of competing imports in year t.

The use of this conventional approach, however, requires that the vector of competitive imports be determined prior to the computation of the vector of gross outputs. Because of this, the authors of Bachue–Philippines have used instead a system of three simultaneous equations that jointly determine import and gross output levels. The first is

$$Z(t) = (I-A)^{-1}F(t) \tag{3.2}$$

where $Z(t)$ is a 13 x 1 vector that represents the *hypothetical* amounts of output that would be produced in year t if there were no competitive imports. The second equation (3.3) relates domestic production in each sector to the hypothetical amount of production that would have occurred in that sector if there were no competitive imports.

$$X_i(t) = \alpha_i(t)\cdot Z_i(t) + [1-\alpha_i(t)]\cdot E_i(t), \qquad i = 1,13 \tag{3.3}$$

where $X_i(t)$ is the level of gross domestic production in sector i in year t; $\alpha_i(t)$ is an import-substitution coefficient, which changes over time at a prescribed rate; $Z_i(t)$ is the hypothetical amount of gross output in sector i in year t that would have occurred if there had been no competing imports; and $E_i(t)$ is the exogenously determined amount of exports for the goods produced in sector i in year t. The third equation in the output determination segment of the model is used to calculate the sectoral levels of imports.

$$Im(t) = F(t)-(I-A)X(t) \tag{3.4}$$

where $Im(t)$ is the 13 x 1 vector of imports in year t and $X(t)$ is the 13 x 1 vector of gross domestic output levels in year t. Although the idea of simultaneously determining import and gross output levels is certainly a good one, the implementation of that idea in the three equations above results in the questionable implication that an increase in the export of output of sector i, *ceteris paribus*, always causes imports of that sector's goods to increase. This can be seen in the following numerical example.

Let us consider the consequences of exogenous one-unit increases in exports of the good produced in sector i. To make the argument concrete, assume that it takes 1.5 units of gross output in sector i to produce 1.0 units of net output. This is equivalent to assuming that the ith element of the diagonal of the inverse of the Leontief matrix, $(I - A)^{-1}$, is 1.5. Now, consider the economic impact of a one-unit increase in $E_i(t)$. Since exports are a component of final demand, $Z_i(t)$ must, according to the assumption above, increase by 1.5 units. If $\alpha(t)$ is 0.5 according to equation (3.3), the increase in gross domestic production must be 1.25 units.[4] There is clearly a problem here. To produce the one additional unit of output requires an increase of 1.5 units in domestic gross output, but only 1.25 units are forthcoming according to equation (3.3). How are the additional 0.25 units obtained? In Bachue–Philippines, it must be through an increase in imports.

This same result can also be demonstrated analytically. For ease of exposition, it is assumed that all the $\alpha_i(t)$ are identical and equal to $\alpha(t)$. Nothing significant in the argument is altered by this assumption. In this case, the expression for the import vector becomes

$$Im(t) = [1 - \alpha(t)] \cdot D(t) + [1 - \alpha(t)] \cdot A \cdot E(t) \qquad (3.5)$$

where $D(t)$ is a 13×1 vector of domestic demand for the outputs of the 13 sectors in year t. In the Bachue–Philippines model $D(t)$ is determined by conditions in year $t - 1$ and $E(t)$ is exogenous. Therefore, it is legitimate to allow $E(t)$ to increase while $D(t)$ is held constant. Clearly, whenever the ith sector's exports rise, its imports must also do so, as must the imports of all other sectors providing intermediate inputs into sector i.

It is possible that increases in exports cause increases in imports under some circumstances. To elevate this notion to a general rule that must be maintained in the long run seems questionable, however. In any case, policy makers doing simulations of various possible export paths should keep in mind the relationship between imports and exports in the Bachue–Philippines model.

It should be noted in passing here that the $\alpha_i(t)$ in equation (3.3) are determined exogenously for the years 1965 through 1975 and are assumed to change at an exogenously predetermined positive rate thereafter.[5] With the passage of time all the $\alpha_i(t)$ approach unity asymptotically. In other words, it is assumed that the Philippines will come to import less and less as a proportion of its hypothetical (without imports) output levels. Thus, import substitution comes about exogenously without any explicit actions on the part of policy makers. This may be an unreasonable assumption in certain contexts, and in those circumstances it should be revised.

Problems also arise in the dynamic assumptions used in the Bachue–Philippines model. The authors provide readers with three choices of dynamic specifications. The simplest is the pure demand model in which there are no supply constraints. The dynamics of this model may be easily summarized. Begin first with a vector of final demands. This is translated into a vector of gross outputs. From that vector the model determines the distribution of personal income in period t and the distribution of consumption expenditures in period $t + 1$.[6] Since the amount and distribution of investment expenditures and government expenditures are essentially exogenous, knowing the distribution of personal consumption expenditures in period $t + 1$ is sufficient to determine fully the vector of final demands. Given this vector a new vector of gross outputs is determined and the process continues.

This formulation is clearly unusual, to say the very least. Output is produced with absolutely no consideration for any factors of production. Thus, the quantities of capital, labor, land, and skills have no impact on the ability of the country to produce output. Further, this formulation makes no allowance for technological progress.[7] This is, of course, in sharp contrast to the approach taken by Kelley, Williamson, and Cheetham, who maintain that biased technological progress is an important element in the story of Japanese economic development. This view that supply factors play no role in the process of development is not a plausible one. It is supplemented in the Bachue model with alternative specifications that allow some, albeit quite limited, role for supply forces.

In the second option, supply factors are introduced by the assumption that gross national product grows at a constant rate each year. If the growth rate of the computed gross national product falls short of the exogenous growth rate, then all elements of aggregate demand are increased so that gross national product grows rapidly enough. On the other hand, if the growth of computed GNP is too rapid, all elements of aggregate demand are reduced proportionally so that output grows at the exogenously given rate. This option is in some dimensions even worse than the specification in which supply does not enter at all. First, since the rate of growth of GNP is predetermined, supply factors still have no influence on the rate of growth of outputs, just as in the original case. One cannot ask about the effect of encouraging capital formation on output growth because in this framework, as in the first one, input growth has no effect on output growth. In the first framework, at least, one could ask questions about the impacts of various policies on the rate of GNP growth. In the second specification, however, nothing the government does can affect the rate of GNP growth. Any policies that affect the rate of population growth will affect the rate of per capita output growth, because the rate of growth of GNP is fixed. This is not a very plausible framework in which to discuss development planning aimed at increasing the rate of output growth. It may have some use in answering questions about the effect of various policies on the distribution of income given that the policies have no effect on growth. Unfortunately, the important questions concerning the trade-offs between inequality and growth cannot be addressed in this version of the model.

Most of the runs and most of the analysis are based on the second version of the model, in which both the rate of output growth and the quantity of investment in each year are taken to be exogenous. In other words, most of the simulations of the Philippine economy assume that output growth and investment are both unrelated to one another and unrelated to anything else in the model. The authors realize that many people consider these assumptions to be unrealistic in the context of a model of long-term economic and demographic change. Therefore, they have performed some sensitivity experiments with variants of the model that allow the rate of economic growth and investment to depend in part on economic and demographic conditions. The demand-dominated model discussed above is one variant of the basic model that is used in these runs. Since supply conditions play no role in this model and investment is still exogenously determined, its usefulness for policy analysis is dubious. A second variant makes the rate of growth of the economy and the level of investment positively related to the balance of payments surplus (or, equivalently, negatively related to the balance of payments deficit). That form of the model is still demand-dominated, but the constraint on growth is at least related to the character of the development process.

The third variant introduces aggregate supply considerations for the first time. Here the rate of growth of aggregate output is determined by the rate of growth generated by a one-sector two-input Cobb–Douglas production function with an exogenously given rate of technological progress. All the capital stocks in the country are aggregated (in an unspecified manner) into a single capital stock. All laborers in the country are aggregated regardless of their wage rates, location, sex, age, and education. Investment is also made endogenous in this variant of the model and depends basically on the rate of growth of aggregate demand lagged one period. Although these supply-side considerations are quite rudimentary, they are a small step in the right direction. The final variant of the model is identical to this one with the exception that the rate of technological progress is positively related to the rate of population growth.

In broad terms, the feature of the Bachue model that most policy makers will have difficulty accepting is the limited role given to supply constraints in the development process. This is not to argue that the process of long run economic and demographic change is to be wholly accounted for in terms of supply-side forces, only that supply- and demand-side considerations interact in an important fashion. The Kelley, Williamson, and Cheetham model of economic development in Japan provides a good example of one way in which the demand and supply sides of the development process can be successfully integrated. Policy makers interested in using the Bachue model may wish to supplement it with some of the ideas implemented there.

There is one important exception to the observation that the supply side of the Bachue model is underdeveloped. This relates to the specification of production possibilities in traditional agriculture. It is assumed that labor productivity in the production of domestic food crops grows at most at a rate $r(t)$ per year. The precise formulation used in the model is

$$X_1(t)_{\max} = L_1(t)_{\text{est}} \cdot \frac{X_1(t-1)}{L_1(t-1)} \cdot [1 + r(t)] \qquad (3.6)$$

where $X_1(t)_{\max}$ is the maximum possible amount of output of domestic foodstuffs in year t, $L_1(t)_{\text{est}}$ is the estimated labor force in the production of domestic food crops in period t, $L_1(t-1)$ is the actual labor force in the production of domestic foodstuffs in period $t-1$, and $r(t)$ is an endogenous but predetermined rate of growth.[8] The labor force in the production of foodstuffs must be estimated from the experience of past years in order to eliminate simultaneity from the model. The equation determining the estimated labor force in the production of domestic foodstuffs in period t is

$$L_1(t)_{est} = L_1(t-1) \cdot \frac{L_1(t-1)}{L_1(t-2)} \tag{3.7}$$

which simply assumes that the current year's rate of increase in the labor force in that sector will be identical to the previous year's rate. Thus, if the rate of growth of the labor force in domestic food production varies from year to year, $r(t)$ may differ somewhat from the *ex post* maximum rate of growth of labor productivity.

If, after the proportional adjustment of all the components of final demand upward or downward to meet the predetermined rate of aggregate output growth, the production of domestic foodstuffs exceeds the maximum output as determined in equation (3.6), there is a response in terms of imports. Gross output of foodstuffs in period t is set equal to $X_1(t)_{max}$ calculated in equation (3.6), and the vector of gross outputs so amended is then used in equation (3.4) to determine a new vector or imports. In this manner it is assumed that imports adjust in the current period to the output limitation.

The relationship between the actual output of foodstuffs and the maximum possible output in each year is assumed to affect the following year's ratio of agricultural to nonagricultural value-added per unit of output in current prices. To understand how this occurs, it is necessary to discuss the process by which sectoral value-addeds per unit of output in current prices are determined. There are thirteen sectors in the model, but the assumption is made that value-added per unit of output in current prices can take on only two values in a given year, one for the four agricultural sectors (domestic food crops, export crops, livestock and fishing, and forestry) and one for the nine nonagricultural sectors (mining, modern consumer goods, traditional consumer goods, other manufacturing, construction, transportation and public utilities, modern services and wholesale trade, traditional services and retail trade, and government services and activities not elsewhere classified).

Before proceeding to a discussion of how the ratio of the two value-addeds changes, it is useful to stop for a moment to evaluate the plausibility that value-added per unit of output in current prices takes on only two values. On the standard assumption that one physical unit of output is that which can be purchased by one currency unit (in this case, by one million Philippine pesos), value-added per unit of output in current prices for 1965 can be determined from the data underlying the input–output used in the model.[9] These figures are given in Table 1. They show that, although value-added per unit of output is generally higher in the agricultural sectors than in the nonagricultural sectors,

TABLE 1 Value-added per unit of output in current prices by sector: Philippines, 1965[a]

Sectors	Value-added per unit of output in current prices
Agricultural	
Domestic food crops	0.907
Export crops	0.910
Livestock and fishing	0.815
Forestry	0.870
Nonagricultural	
Mining	0.733
Modern consumer goods	0.636
Traditional consumer goods	0.570
Other manufacturing	0.620
Construction	0.650
Transportation and public utilities	0.712
Modern services and wholesale trade	0.831
Traditional services and retail trade	0.765
Government services and activities not elsewhere classified	0.985

[a] Data from: Rodgers *et al.* (1976), pp. IV-17 and IV-18.

constancy is not well approximated. Below, an improved procedure is discussed that makes use of the figures in Table 1.

Let $v_a(t)$ be the single value-added per unit of output in current prices in the agricultural sectors in year t and $v_n(t)$ be the single value-added per unit of output in current prices in the nonagricultural sectors in year t. The ratio of the agricultural value-added to the nonagricultural value-added is given by

$$\frac{v_a(t)}{v_n(t)} = \frac{v_a(t-1)}{v_n(t-1)} + k_1 \left[\frac{X_1^*(t-1) - X_1(t-1)_{\max}}{X_1^*(t-1)} \right] \tag{3.8}$$

where $X_1^*(t-1)$ is the amount of output of the domestic foodstuffs sector in period $t-1$ after any proportional adjustments in the elements of aggregate demand but before the application of the productivity limit and k_1 is a constant that is set equal to unity in the simulations.

Several aspects of this specification deserve comment here. First, equation (3.8) relates changes in a value-added ratio to the excess demand or supply for domestic foodstuffs. A much more natural formulation would use the excess demand or supply of domestic foodstuffs to influence the relative price of domestic foodstuffs. Second, changes in the value-added ratio are assumed to be influenced only by the relation between the supply and demand for foodstuffs. Supplies and demands for other goods are assumed to have no impact. Third, the hypothesis that k remains constant at unity is quite weak, particularly for such an important link in the argument. There is no empirical evidence to suggest that k is either constant or in the vicinity of unity. Finally, it is not clear that value-added per unit of output in current prices in domestic foodstuffs and export crops changes proportionally, since the price of the latter can be expected to be closely aligned to world prices.

Given the ratio of prices determined in equation (3.8), the level of prices is determined as follows

$$\sum_{i=1}^{4} v_a(t)S_i(t) + \sum_{i=5}^{13} v_n(t)S_i(t) = 1 \qquad (3.9)$$

where $S_i(t)$ is ratio of value-added in constant prices in sector i in year t to aggregate output (in current prices) in that year, and where the sectors numbered 1 through 4 are the agricultural sectors and those numbered 5 through 13 are the nonagricultural sectors.

This process of deflating value-added per unit of output in equation (3.9) is quite unusual. To understand the problems with equation (3.9) requires some preparation. In a model of the kind we are considering there is a relationship between the input–output coefficients, sectoral value-added per unit of output in current prices, and output prices. That relationship is

$$(I - A')P(t) = v(t) \qquad (3.10)$$

where $P(t)$ is the 13×1 vector of output prices in year t, A' is the transpose of the input–output matrix, and $v(t)$ is the 13×1 vector of value-added per unit of output in current prices in year t.

Given the standard assumption that physical units of output are defined to be a quantity whose value is worth one currency unit (one million Philippine pesos, in this case), all the output prices in the base year are unity. Given those base-year prices, equation (3.10) can be used to obtain the value-addeds per unit of output in current prices shown in Table 1. As was discussed above, however, those are not the value-added figures used in the base year. Instead, the authors of Bachue–Philippines utilize their bi-level value-addeds derived from equations (3.8) and (3.9) to determine current output prices as follows:

$$P(t) = (I - A')^{-1} \cdot v(t) \qquad (3.11)$$

This method of price determination is seriously deficient as used in the model. First, if the correct value-addeds for 1965 were used without the level modification in equation (3.9), the current prices in 1965 would all be unity. Equation (3.9), however, raises all the value-addeds by some proportion and all the output prices by the same proportion. If the other equations in the model appropriately take the nonunitary prices into account (which is shown below not to be the case) this procedure is technically correct. When the bi-level value-addeds are used, however, the prices for 1965 without level adjustment are no longer equal to one another or to unity, as is, of course, also the situation after the level change.

The input–output coefficients in the matrix A, though, are computed for 1965 on the assumption that all the output prices are identical. Thus, the base-year prices, value-addeds, and input–output coefficients are inconsistent with one another. This is an important problem, and one that, because of equation (3.8), affects other years as well.

Two further problems affect the price system in Bachue–Philippines. First, because prices are not all unity, a distinction has to be made between expenditures in currency units and quantities of goods purchased. Unfortunately, this is not done in the model. The implicit assumption that output prices are indeed unity pervades much of the model. The result of this is that quantities are generally computed incorrectly. The second problem concerns the income determination segment of the model where an improper deflation causes the income flows to be mismeasured.

Any policy maker interested in using Bachue seriously must correct these problems. The simplest set of corrections to make in the spirit of the Bachue model are, first, to use the value-added per unit of output data from the 1965 input–output table. Next, keeping the within-agricultural and within-nonagricultural relative prices constant, modify the agricultural and nonagricultural price ratio as in equation (3.8). Third, use the new price vector computed in each year to determine value-added per unit of output in each sector in that year by means of equation (3.10). Fourth, use the vector of value-added per unit of output computed in step three with the appropriate base-year figures in the income distribution calculations. This four-step process will ensure that the price, value-added, and income distribution figures used in the model are, at least, consistent.

3.2 DETERMINATION OF THE COMPONENTS OF FINAL DEMAND AND SAVINGS

In all versions of the model except the pure demand-driven case, each component of final demand is computed twice. Generally, the initial values of the final demands for the 13 sectoral outputs are inconsistent with the predetermined level of aggregate output. To avoid this inconsistency and to maintain the predetermined level of output, the final demands for the output of the 13 sectors are altered proportionally. In the discussion below, we treat only the *ex ante* or first-stage values of the components of final demand.

Consumption and Savings

One of the most interesting features of the Bachue model is the treatment of the distribution of income. Household income in the urban and rural areas are divided into deciles, and savings and consumption expenditures are determined separately for each of them. Average household consumption of the output of sector i by households in the dth decile of the income distribution in location k in year t is given by

$$C^*_{idk}(t) = \theta_{id}(t) \cdot \{\alpha_{ik} + \beta_{ik}[Y^*_{dk}(t)_{est} - S^*_{dk}(t) - T^*_{dk}(t)] + \gamma_{ik} \cdot A^*_{dk}(t) + \delta_{ik} \cdot C^*_{dk}(t)\} \quad (3.12)$$

where $C^*_{idk}(t)$ is the average household consumption of the output of sector i in year t by households in location k in the dth decile of the income distribution; $\theta_{id}(t)$ is a multiplicative factor relating to prices[10]; $Y^*_{dk}(t)_{est}$ is the estimated average income in year t of households in location k who are in the dth decile of the income distribution[11]; $S^*_{dk}(t)$ is the average household savings accumulated in year t by households in location k who are in the dth decile of the income distribution; $T^*_{dk}(t)$ is the average level of income taxes paid in year t by households in location k in the dth decile of the income distribution; $A^*_{dk}(t)$ is the mean number of adults in location k in year t who live in households in the dth decile of the income distribution; $C^*_{dk}(t)$ is the mean number of children in location k in year t who live in households in the dth decile of the income distribution; and α_{ik}, β_{ik}, γ_{ik}, and δ_{ik} are sector- and location-specific constants. The $\theta_{id}(t)$ in equation (3.12) are defined as follows

$$\theta_{id}(t) = \frac{Z^*_d(t)}{\left\{\left[\frac{P_i(t)}{P_i(0)} - 1\right]\epsilon_i + 1\right\}}, \qquad i = 1, \ldots, 13 \quad (3.13)$$

where $Z^*_d(t)$ is a factor that depends upon all the variables on the right-hand side of equation (3.12), the $P_i(t)$, and the ϵ_i;[12] $P_i(t)$ is an element of the $P(t)$ vector derived from equation (3.11) above; and ϵ_i is a sector-specific constant.

There are several aspects of this specification that require comment. First, the prices used should be from a procedure such as that outlined at the end of section 3.1 above. Second, equation (3.13) is not specified in terms of relative prices, but in terms of the level of a single price. A preferable manner of incorporating prices into demand functions is to use a known system of demand equations such as the Stone–Geary demand structure used in the Kelley, Williamson, and Cheetham model. We shall say more about this below. Third, the denominator in equation (3.13) may over the course of a long simulation period come to approach zero for some goods, causing the resulting pattern of consumption expenditures to become implausible. Fourth, the term $[Y^*_{dk}(t)_{est} - S^*_{dk}(t) - T^*_{dk}(t)]$ is supposed to equal the average consumption in year t of households in location k in the dth decile of the income distribution. The implicit assumption made here is that taxation has no effect on savings and affects only consumption. This assumption may not be true in many cases. Policy makers who wish to use the Bachue model to analyze policies involving increases or decreases in income taxes should ascertain first whether this particular assumption is appropriate for their countries.

Two important points concerning the consumption specification involve aggregation. First, aggregating across commodities within income deciles, the following equation must obtain:

$$\sum_{i=1}^{13} C^*_{idk}(t) = Y^*_{dk}(t)_{\text{est}} - S^*_{dk}(t) - T^*_{dk}(t) \tag{3.14}$$

In words, the sum of the average expenditures on all goods in year t by households in location k in the dth decile of the income distribution must equal their average total consumption expenditures. Unfortunately, holding $Z^*_d(t)$ constant and altering any variable on the right-hand side of equation (3.12) will, in general, falsify equation (3.14). The sum of the average expenditures on all goods will no longer equal average total consumption expenditures. This problem must be resolved somehow, and it is in this context that $Z^*_d(t)$ plays a role in the consumption specification. Every time anything affecting consumption changes, $Z^*_d(t)$ and therefore the $\theta_{id}(t)$ move up or down until equation (3.14) is satisfied once more. This could easily lead to quite peculiar results. Suppose, for example, that a certain γ_{ik} is positive. One might think that this implies that when $A^*_{dk}(t)$ rises, $C^*_{idk}(t)$ rises, but this is not necessarily the case. The adjustment factor $\theta_{id}(t)$ may fall sufficiently under some circumstances as a result of the increase in $A^*_{dk}(t)$ that $C^*_{idk}(t)$ will actually fall. Such problems make equation (3.12) a very poor specification of the relationship between consumption levels, incomes, and prices. The weakness of this formulation should not be viewed as the inevitable result of the inherent complexity of the problem. There is a substantial literature on systems of demand functions that aggregate correctly and in which price and income elasticities enter in a consistent and coherent manner. Indeed, in the earliest of the models reviewed here, the Kelley, Williamson, and Cheetham model, such a system is used. For a discussion of those equations see section 7.2 below. It may be useful in future work on the Bachue model to replace the set of consumption equations with a set that has more plausible properties.

The second point regarding aggregation concerns aggregation across income deciles. The object of the consumption specification is to determine the total consumption demand for the outputs of each of the 13 sectors. This can be done by aggregating across income deciles and then summing across locations. It is instructive to note in this regard that none of the parameters in equation (3.12) except the correction factor $\theta_{id}(t)$ depends upon the decile level in the income distribution. If $\theta_{id}(t)$ were totally independent of the decile level, a simple summation across deciles in a particular location would yield a relationship in which total consumption of good i in location k would depend linearly upon total household income in location k, total savings by households in location k, and total taxes paid by households in location k. In other words, were it not for the unusual formulation in which consumption expenditures require the proportional adjustment described above, disaggregation by income level would be irrelevant for the determination of total consumption levels, except to the extent that such a disaggregation is required to compute total savings or total income taxes paid. Indeed, because of this, it is not surprising to learn that the effect of changes in the income distribution on the other endogenous variables in the model is quite small (see Rodgers *et al.* 1976, p. VII.9 and VII.10).

Before leaving the subject of consumption, it is important to make note of an equation that does not appear in the model, one relating consumption expenditures to the number of units consumed. The absence of this equation implies that output prices are thought to be unity. It was shown above, however, that this is not the case. Those interested in using the Bachue-Philippines model should supply the missing equations.

Average household savings in year t by households in location k that are in the dth decile of the income distribution is given by

$$S^{*}_{dk}(t) = \begin{cases} \alpha_k + \beta_k Y^{*}_{dk}(t)_{\text{est}} + \gamma_k [A^{*}_{dk}(t) + C^{*}_{dk}(t)] & d = 6, 10 \\ 0 & d = 1, 5 \end{cases} \tag{3.15}$$

where the variables are all as defined above in equation (3.15), but α_k, β_k, and γ_k are different constants. There are two important aspects of this savings function to note. First, it is discontinuous. Households in the lower five deciles of the income distribution are assumed not to save anything.[13] Second, the parameters of the savings function are independent of the decile level in the income distribution. Aggregating over the upper five income deciles implies that total household savings is a linear function of the total amount of income earned by households with incomes above the median, the total number of people who live in the households, and the total number of such households. Thus, for the purpose of computing total household savings, discrimination between two income groups is all that is necessary.

Before, we leave the topic of savings, one further set of remarks is in order. There is no direct connection between savings and investment in the Bachue model. In most of the simulation runs, *ex ante* investment is exogenous and thus its magnitude is independent of the amount saved. There is a weak indirect connection between savings and investment in those runs where output growth is predetermined. Increasing savings implies, holding everything else constant, that consumption will fall. If before the increase in savings aggregate demand was equal to aggregate supply, after the increase aggregate demand would be too small and each element of final demand, including consumption and investment, would have to be proportionally increased so that the equality could be maintained. In this manner, changes in savings can have a small impact on levels of investment. In most runs, however, this route for savings to affect the economy is attenuated even further by the assumption that output growth is unaffected by the growth of the capital stocks. The authors reveal on page IV.24 (footnote 1) that model outcomes are insensitive to household savings. The discussion here makes it evident why this is the case. Policy makers wishing to use the Bachue framework ought to consider whether the connections between investment and savings and between capital stocks and outputs are appropriate for their countries. If they are not appropriate, the policy makers may want to consider some of the alternative specifications of these relationships used in the models reviewed here.

Investment

Ex ante investment in Bachue is considered to be exogenous in most of the simulation runs, independent both of savings and the rate of output growth. For the period after 1975, it is assumed that *ex ante* total investment grows at 7 percent per year. Investment, it should be recalled, does not play a significant role in the Bachue model because capital is generally not treated as a factor of production. The capital stocks in various sectors do, however, play a small role in determining the distribution of incomes and the distribution of gross outputs. Two aspects of investment are relevant here. The first concerns the 13 x 1 vector of investment expenditures by sector of production. Actually, there are three such vectors in the model, one for government investment, one for investment in dwellings, and one for other private investment. The total amount of government investment is given as a fixed exogenous fraction of the exogenous amount of total investment. The total amount of investment in dwellings is endogenous, depending on the share of rents in total household consumption. The total amount of other investment is taken as a residual maintaining the exogenous amount of *ex ante* investment. Government investment and other investment totals are allocated to sectors according to fixed exogenous proportions. All investment in dwellings is allocated to the construction sector.

The second aspect of investment to be discussed is the allocation of capital according to sector of application. This is done through the use of a set of fixed incremental capital-output ratios. Clearly, the amount of investment required on the basis of those ratios may not equal the exogenous amount of investment funds available. To resolve this inconsistency, all the incremental capital-output ratios are proportionally increased or decreased so that the amount invested is equal to the amount available for investment. Thus, although the incremental capital-output ratios are nominally fixed, the *ex post* incremental capital-output ratios can vary considerably from year to year. Thus, a shortage of capital can never affect the rate of growth of output or the character of the development process. Fortunately, the allocation of capital among the various using sectors has little impact on the other facets of the model.

As with consumption expenditures, investment expenditures are not deflated before they are added to final demand. This should be modified by users of the Bachue-Philippines model.

Government Expenditures

There are several possible ways of treating government expenditures in the model. The most interesting alternative was the one used in the base run. There government expenditures were assumed to be determined by the following equation:

$$G(t) = a \cdot GDP(t-1) \cdot Pop(t)^{0.34} + U(t) \qquad (3.16)$$

where $G(t)$ is the amount of government expenditures in period t, a is a constant, $GDP(t-1)$ is gross domestic product in period $t-1$, $Pop(t)$ is the total population of the country in period t, and $U(t)$ are additional expenditures on programs like education and public works. It is interesting to note with regard to this specification that the ratio of $G(t)$ to $GDP(t-1)$ is a positive function of population and the share of those additional expenditures in *GDP*. Thus, even if the latter is constant, the share of government expenditure in *GDP* is assumed to grow over time. Policy makers who are not in a situation in which it is reasonable to expect such an evolution should make appropriate modifications to this specification before they use the model.

Government investment is a fixed fraction of the exogenously determined amount of total investment. What remains of total governmental expenditures is called government consumption. Government consumption is allocated to sectors according to a fixed set of coefficients. Thus if 10 percent of government consumption is spent on domestic foodstuffs in 1965, 10 percent of government consumption will be spent on domestic foodstuffs in 2005. It is mildly curious that the allocation of these expenditures appears to have nothing to do with the quantity and the nature of the expenditures under the category $U(t)$. Thus, for example, increasing the amount of educational expenditures reflected in the $U(t)$ variable will not alter the *allocation* of total government expenditures by sector.

Like other elements of final demand, government expenditures are inappropriately undeflated.

Exports

The final element in final demand is exports. It is assumed, for the years following 1969, that exports in each sector grow at an exogenously given sector-specific rate.

3.3 THE DISTRIBUTION OF INCOME AND EMPLOYMENT

The Bachue–Philippines model exceeds all the other second-generation models reviewed here in the detail and care used in describing the labor market and the distribution of income. As we discussed above, neither the labor market nor the distribution of income has much direct impact on aggregate economic phenomena. For example, in the main version of the model, a more rapidly growing labor force cannot induce more rapid output growth because the rate of growth of aggregate output is considered to be exogenously determined. Nor can a rapidly growing labor force cause profits and therefore investment to increase because, first, profits are not directly related to either the number of workers or their wage rates and, second, total investment is independent of profits (and everything else in the model). Nonetheless, it is still of some interest to ascertain what, if anything, can be said about the effects of various policies on the distribution of income and employment. The question remains open, however, concerning the trade-off between growth and income distribution. Certain policies may worsen inequality for some period of time, but make everyone better off in the long run. Such policies cannot be studied in the context of the Bachue model.[14] Rather, what can be studied are the distributional aspects of some policies abstracting from any impacts they might have on the pace of development.

Labor Force Participation Rates

The Bachue model determines 176 labor force participation rates in each year. For the purpose of computing these rates, people are divided according to the following characteristics: sex (two categories), marital status (two categories – married or not for females and household head or not for males), education (two categories), age (eleven 5-year age groups) and location (two categories). The labor force participation rates for male household heads are assumed to remain constant at their 1965 levels. The remaining labor force participation rates are endogenously determined.

Since the variables that enter the labor force participation rate equations pertaining to groups of workers have a substantial overlap, we shall focus our attention here on the nature of those variables, rather than on the more numerous individual equations. One variable is the proportion in the previous period of the total number of employed people in a given location who work in modern sectors. It is assumed in the model that as this rate rises the labor force participation rates increase for all groups except single females who are above the age of 25 and who live in urban areas. Examples of the meanings of this specification are easy enough to cite. It says, among other things, that as production in the rural areas shifts away from food production and moves toward forestry, construction, and transportation and public utilities, the labor force participation rates of married females will rise. This seems to presuppose that there is something about forestry, construction, and transport which induces rural married women to participate more readily in these sectors than they do in traditional agriculture. Whether the sign of this effect is correct must, it seems to me, remain open to serious question. Similarly, the model assumes that the labor force participation rate of urban male non-household heads also rises when the share of total urban employment that is in traditional pursuits diminishes. Yet a substantial number of urban male non-household heads are surely relatively young men (or boys) living at home who could more easily participate in the traditional than in the modern sector.

There is a methodological reason to suspect that many of the postulated directions of effect in the labor force participation equations are incorrect. The signs were not derived by regressing the indicated variables on participation rates, but rather by regressing some other variables on the participation rates and assuming that the coefficients remain basically unchanged when the indicated variable is substituted for the one used in the regression. For example, the effects of the proportion in the previous period of the total number of workers in a given locale who are employed in modern pursuits on labor force participation rates were not determined by a regression in which that variable actually appeared. The corresponding variable in the regression is the "percentage of people in the region of residence . . . who were enrolled in school last year and are working in a modern sector this year" (p. V.8). The relationship between the variable in the regression analysis and the variable in the labor force participation rate equations is sufficiently tenuous that it would not be surprising if a number of the signs in the latter equations are incorrect. This procedure of computing regression coefficients used in the labor force participation equations from independent variables that do not appear in those equations is replicated for two other variables.

The second variable used to explain labor force participation rates in location k in period t is the ratio of the arithmetic mean of disposable income in location k in period $t - 1$ to the harmonic mean of the average incomes in the ten income deciles[15] in period $t - 1$. Roughly speaking, that ratio is positively related to income inequality. Since this variable contributes positively to the labor force participation rate, income inequality is positively related to labor force participation rates. If output were allowed to be affected by employment, this relation would play a role in the trade-off between growth and income equality.

The third type of variable that is included in the explanation of labor force participation rates is a set of three location-specific employment shares: (a) the share of employment in construction, transportation, and public utilities, (b) the share of employment in modern services, wholesale trade, and government, and (c) the share of employment in the production of traditional consumer goods, traditional services, and retail trade. It is not worthwhile to detail all the assumptions relating these three shares to the labor force participation rates of various groups. Instead, as an example it will suffice to show the assumptions made with regard to the third share. This share is negatively related to the labor force participation rate of urban male non-household heads below the age of 34, but positively related to their labor force participation rates at higher ages. For rural male non-household heads that share is negatively related to labor force participation rates at all ages. For urban married women that share is positively related to labor force participation rates at all ages, but for their unmarried sisters living in urban areas it is negatively related to labor force participation rates. For all rural females, however, that share is positively related to labor force participation rates (except for unmarried females 15–19 years old, where there is no effect).

The final included variable pertains only to married women. It is

$$\left[\frac{Z_a}{MF_k(a) + 1}\right]^{-1} \tag{3.17}$$

where Z_a is an age-specific constant and $MF_k(a)$ is the marital fertility rate for women of age a in location k. This variable is positively related to age-specific fertility rates and is therefore assumed to be negatively related to labor force participation rates of those women.

It should be noted here that labor force participation rates are assumed not to be influenced either by wage rates or by prices. Making labor force participation rates endogenous is a difficult task. The authors of Bachue should be commended for their efforts in this regard even if the resulting specifications leave room for improvement.

The Determination of Aggregate Levels of Employment and Unemployment

By far the most articulated portion of the Bachue model relates to employment and the distribution of income. Because the urban and rural specifications of the functions in this portion of the model are quite similar, undue repetition will be avoided by focusing solely on the formulations relating to the rural area.

The determination of rural employment and unemployment begins in any year with the predetermined size of the labor force and the number of rural households.[16] These figures are affected over time by rural-to-urban migration (or the reverse), but are assumed to be unaffected by events in the current year. Employment is not computed from consideration of the demand and supply of rural workers, but rather from consideration of the relative wage rates in the various rural sectors in the previous year. The computation proceeds in two steps. First, the employment for period t is computed on the basis of relative wage rates in the traditional and modern sectors in the previous year. Next, the estimate of employment for period t is recomputed by averaging the initial estimate of employment and the level of employment in the preceding year.

More specifically, the expressions used in the model are

$$E_r^*(t) = \alpha \cdot \left[\frac{W_1(t-1)}{W_2(t-1)}\right]^{\beta} \cdot L_r(t) \tag{3.18}$$

$$E_r(t) = (0.5) \cdot E_r^*(t) + (0.5) \cdot E_r(t-1) \tag{3.19}$$

$$U_r(t) = \frac{[L_r(t) - E_r(t)]}{L_r(t)} \tag{3.20}$$

where $E_r^*(t)$ is the first estimate of employment in the rural area in period t, $E_r(t)$ is the final estimate of employment in the rural area in period t, $W_1(t-1)$

is the wage rate in the traditional rural sectors in the year $t - 1$, $W_2(t - 1)$ is the wage rate in the modern rural sectors in the year $t - 1$, $L_r(t)$ is the labor force in the rural area in period t, $U_r(t)$ is the unemployment rate in the rural area in period t, and α and β are positive constants.

The question that immediately arises concerns the meaning of those equations. One possible interpretation would be that the labor force measures the number of people who are willing to work at the prevailing wage rates and therefore provides the rural economy with a supply-of-labor curve of infinite elasticity up to $L_r(t)$. Employment then would be determined by demand conditions. But it is not clear under this interpretation why the *demand* for rural labor should be positively associated with the wage rate in the traditional rural sectors, although it seems plausible enough to assume that it is negatively associated with the wage rate in the modern rural sectors. An alternative interpretation is that the demand for rural labor is infinitely elastic. In this circumstance employment is determined by the supply of labor. This requires a new interpretation of $L_r(t)$, however. It would now be the labor force that would be employed at some very high wage. If wages were not sufficiently high, some members of the potential labor force would not work and therefore employment would be reduced. Under this interpretation both the wage rate in the traditional sector and that in the modern sectors should be positively related to employment.

Clearly, equation (3.18) is a mixed case. The implicit assumption seems to be that, with regard to the traditional rural employment, the supply-side effects dominate and, with regard to the modern rural employment, demand-side effects dominate. This is certainly possible. Still, if that is the story the authors wish to tell, it would have been preferable to weight the effects of wage rate changes according to the relative numbers of people employed in the modern and traditional sectors. For example, if modern rural employment accounted for only 1 percent of total rural employment, then a 1 percent increase in the traditional wage rate may possibly have quite a different effect on employment than a 1 percent decrease in the wage in the modern sectors. At present, in equation (3.18), it is assumed that the effects on employment of those two wage changes are identical.

Value-Added Shares

One important step in the process of determining the income distribution in the Bachue model is the division of the total value-added in each sector into a labor and a nonlabor share. In the rural sector this division is done basically by assumption. In all those sectors except one, it is assumed that the share of value-added in constant prices remains forever at the level observed in 1965. With regard to rural transportation and public utilities a different approach is used. The nonlabor share of value-added in that sector is assumed to be a linear function of the percentage of urban modern value-added in total value-added all measured in constant prices. The coefficients of that linear relation are derived from observations in the Philippines in 1965 and several developed countries (particularly Japan) around 1960.

Three points deserve brief mention here. First, as shall be shown below, what follows in the Bachue model requires that the value-added share assumptions be applied to value-added measured in current prices, not constant prices. Second, determining value-added shares as linear functions of the percentage of urban modern value-added in total value-added is extremely restrictive. It gives essentially no scope for short run policies to operate by changing value-added shares. And this leads to the third observation. Valued-added shares are an important determinant of the income distribution. Such a weak specification of how they behave is not consistent with the thrust of the modeling effort. Policy makers interested in using the Bachue framework should certainly pay some attention to improving the assumptions made in this portion of the model.

Distribution of Employment

Bachue distinguishes between two sorts of employment, self-employment and wage employment. Self-employment in rural modern sectors is given by the following equation:

$$E_{si}(t) = f_i \cdot \frac{W_{ni}(t-1)}{[W_{si}(t-1)]} \cdot H_r(t) \tag{3.21}$$

where $E_{si}(t)$ is the number of people self-employed in the ith rural modern sector in year t; $f_i\{[W_{ni}(t-1)]/[W_{si}(t-1)]\}$ is a sector-specific function whose value is negatively related to the value of its argument, the ratio of average wage income $W_{ni}(t-1)$ to average nonwage income $W_{si}(t-1)$ in the ith rural modern sector in year $t-1$; and where $H_r(t)$ is the number of rural households in year t. The assumption made here is that as wage income increases relative to nonwage income, the number of nonwage income earners decreases. Suppose for a moment we apply this assumption to a hypothetical example in a particular rural modern sector – forestry. For the sake of discussion, let there be two types of forestry workers, those who chop down trees for themselves (self-employment) and those who do the same task for a company (wage-employment), and let their incomes be initially identical. Now, let the income paid to those with wage employment increase exogenously. A demand-side interpretation would be that workers would tend to move into the now higher-paying wage employment and out of self-employment. This is, of course, what is predicted by equation (3.21). A supply-side interpretation would be that the company would hire fewer loggers and thus cause self-employment to rise. This is, of course, the opposite of what is predicted by equation (3.21). But does the demand-side effect *always* dominate in this context? The answer to this is certainly unclear, but the question can provide some guidance to those who may wish to improve upon the specification in equation (3.21).

Wage employment in the rural modern sectors is computed using the following equation:

$$E_{ni}(t) = [V_i(t) \cdot K_i(t)^{-\beta_i} \alpha_i^{-1}]^{1/\gamma_i} e^{\delta_i t} - M_i(t) \tag{3.22}$$

where $E_{ni}(t)$ is the number of people working for wages in the ith rural modern sector in year t, $V_i(t)$ is the value-added in the ith rural modern sector in year t measured in constant prices, $K_i(t)$ is the capital stock in the ith rural modern sector in year t, $M_i(t)$ is the number of wage laborers who leave (enter) the modern rural sectors and take (leave) employment in the export crop sector,[17] β_i is the self-employment share of value-added in rural modern sector i, γ_i is labor's share in value-added, δ_i is a parameter relating to technical progress, and α_i is a constant.

There are several facets of this equation that deserve mention. First, there are two terms on the right-hand side of the equation, one representing the amount of labor that would be required to produce the appropriate amount of value-added if a Cobb–Douglas production function is appropriate. Given the assumption that value-added shares are constant in any given period, this choice seems to be the correct one. After the appropriate employment level is determined, however, a factor is added to that number – the number of people who formerly held jobs in the export crop sector but who will be employed in the modern rural sector in the current year (or the reverse if migration is toward the export crop sector). At first glance this seems rather unusual. If the Cobb–Douglas production function is indeed appropriate, then why should anything be added to the employment figure it generates? The people who come to be employed in the modern rural sectors should be a portion of that total, not added on to that total! The specification in equation (3.22) seems on its face to be roughly analogous to determining the temperature using the following approach. First, find an accurate thermometer and give it adequate time to measure the temperature correctly. Read the thermometer and take as your estimate of the temperature the reading on the thermometer plus or minus some other figure, such as the humidity or the rainfall within the last month.

Although accurate, this characterization is somewhat unfair. We are not dealing here with a neoclassical model of the economy and the distribution of income, but rather with a model that incorporates a number of assumptions that would not be included in such a model. As the authors correctly perceive, solving for employment using the inverted Cobb–Douglas production function would produce wage rate differentials that are terribly unrealistic. Therefore, in order to keep the wage rate differentials within a plausible range while maintaining all the other assumptions made in the model, the authors are forced to modify the Cobb–Douglas production function as they have done. Thus, given the other assumptions in the model, the specification in equation (3.22) may be preferable to the more obvious one in which just the inverted Cobb–Douglas production function is used. To maintain roughly the same story as in Bachue and to allow the inverted Cobb-Douglas production function to determine employment would require that the outputs of the rural modern sectors have different prices. This is not implausible, and individuals wishing to use the Bachue model in the future may wish to compute these relative prices.

Self-employment and wage employment in the export crop sector are separately determined. Wage employment in the export crop sector is computed by dividing labor's share of value added in that sector in *constant prices* by an estimate of the annual income of those employed in that sector. There are two problems with this approach. First, employment cannot properly be determined in that manner. The correct procedure would be to divide labor's share in value-added in *current prices* by the estimate of the average annual income of employees. The second problem relates to the estimate of average annual income. Instead of discussing this difficulty here, however, we will treat it below as one aspect of a more general problem.

Self-employment is calculated using the following equation:

$$E_{se}(t) = \frac{V_{se}(t)}{W_{se}(t)} - M_{se}^{*}(t) + M_{se}^{**}(t) \tag{3.23}$$

where $E_{se}(t)$ is self-employment in the export crop sector in year t, $V_{se}(t)$ is value-added in that sector in constant prices in year t, $W_{se}(t)$ is an estimate of the average annual income of self-employed people in the export crop sector in year t, $M_{se}^{*}(t)$ is the number of self-employed people in the export crop sector in the previous period who are employed in the modern rural sectors in the current period, and $M_{se}^{**}(t)$ is the number of self-employed people in the export crop sector in the current period who worked in the previous period in traditional agricultural pursuits. As was indicated in the preceding paragraph, the correct way to compute employment would be to divide value-added in current prices by an estimate of the average annual income of self-employed people. As employment is calculated in equation (3.23), it will not, in general, equal that figure. It might be argued that the approach taken in equation (3.23) is an alternative to the one suggested above. If that suggestion is to be taken seriously, policy makers interested in using the Bachue model should take care to ensure that the implications for the labor and nonlabor shares of value-added in current prices are acceptable.

Employment in traditional agricultural production is derived as a residual after total rural employment and employment in each of the other rural sectors is obtained.

Distribution of Income

Average annual income estimates in the Bachue model are generally computed incorrectly. They are calculated either directly or indirectly from the following equation:

$$W_{ri}(t) = \frac{V_{ri}(t)}{E_{ri}(t)} \cdot v_i(t) \qquad (3.24)$$

where $W_{ri}(t)$ is the average annual income of type r (wage income or self-employment income) in sector i in year t, $V_{ri}(t)$ is the amount of value-added of type r in sector i in year t measured in constant prices, $E_{ri}(t)$ is employment of type r in sector i in year t, and $v_i(t)$ is the value-added per unit of output in sector i in year t in current prices. The correct computation of average annual incomes is accomplished by dividing the appropriate value-added in current prices by the corresponding employment figure. Equation (3.24) is in error because, generally, the appropriate value-added in current prices is not equal to corresponding value-added in constant prices multiplied by value-added per unit in current prices.[18]

Thus, the Bachue model has difficulties both in the determination of the distribution of employment between sectors and in the determination of average annual incomes in each of the sectors. Under some circumstances these problems may be serious, while in others they may be trivial. To be on the safe side, policy makers who are interested in the Bachue framework should correct those problems before trying to obtain meaningful simulation results for their countries.

The data on the distribution of employment and on average earnings in the various sectors are the major inputs into the portion of the model that determines the overall distribution of personal income. That segment is one of the most innovative features of the Bachue model. Rather than describing its entire structure here, we will discuss only the broad outlines of the income distribution determination. Separate distributions for rural and urban households income are computed. It is assumed that both distributions are lognormal and therefore are completely described by two parameters and the mean variance. The means of the two distributions are readily computed given the value-added shares discussed above and the number of households in each of the two areas.[19]

The variances of the two distributions are much more difficult to obtain. First, means and variances of incomes for people employed in the various sectors must be transformed into means and variances of incomes of households where the head is employed in given sectors, and second, the latter figures must be used to determine the appropriate overall variance. As one might imagine, a large number of assumptions are required to go from the data in the model to the variance of the household income distribution. For example, it is posited that in a given sector the incomes of heads and nonheads of households are identical. It is difficult to evaluate a system that is based on such assumptions. Although the assumptions may be technically incorrect, the resulting distributions of income may be good approximations. On the other hand, however, circumstances could arise in which the Bachue procedures may yield poor approximations to income distributions. Policy makers who are interested in using the income distribution feature of the Bachue model should carefully test it on their own data before accepting it as being useful for them.

3.4 THE DEMOGRAPHIC SEGMENT

The Demographic Accounting

The population in the Bachue model is subdivided along four major axes: (a) age (0–1, 1–4, 5–9, 10–14, . . . , 60–64, 65 and over), (b) sex, (c) location (rural and urban), and (d) education (less than primary, at least primary but less than secondary completed, secondary completed or more). The people in each of the categories must be followed across space and through time. The procedures for doing so are well known in the demographic literature. The use of 5-year age groups, however, creates problems because all the single-year data on birth, marriage, and education cohorts are lost. In a footnote on page VI.6 the authors suggest that future versions of the model would be simplified if single-year-of-age acounting were utilized. This is certainly the case, and policy makers interested in adapting Bachue for their own use should take this suggestion of the authors' seriously.

The Determination of the Number of Households in the Urban and Rural Areas

The number of households in the rural and urban areas are determined by applying exogenous rates to population groups disaggregated by age, sex, and location. The authors realize that economic development may change the propensities of various groups to form households, but they have no way to treat this complex phenomenon. Given the inadequate amount of information available, the assumption of constant headship rates may be about as good as any assumption one could make at present. Future work, however, could possibly take adavantage of the fact that the Bachue model also includes marriage rates.

The Determinants of Average Household Size and Composition

The average household sizes in the urban and rural areas are determined simply enough. The total number of people living in each location is divided by the total number of household heads living in each place. The mode of determination of the latter figure is given in the immediately preceding section. The next aspect of the model that requires computation is the composition of households across varying levels of household income. This is accomplished by use of the following three equations:

$$A_{kd}(t) = A_{kd}(0) \cdot \left[\frac{p_{15^+,k}(t) \cdot p_{15^-,k}(0)}{p_{15^-,k}(t) \cdot p_{15^+,k}(0)}\right] \cdot \frac{F_k(t)}{F_k(0)} \tag{3.25}$$

$$C_{kd}(t) = C_{kd}(0) \cdot \left[\frac{p_{15^-,k}(t) \cdot p_{15^+,k}(0)}{p_{15^+,k}(t) \cdot p_{15^-,k}(0)}\right] \cdot \frac{F_k(t)}{F_k(0)} \tag{3.26}$$

"adjusted so that":

$$0.1 \cdot \sum_{d=1}^{10} [A_{kd}(t) + C_{kd}(t)] = F_k(t) \tag{3.27}$$

where $A_{kd}(t)$ is the average number of adults in households in the dth decile of the income distribution in location k in year t, $C_{kd}(t)$ is the average number of children in households in the dth decile of the income distribution in location k in year t, $p_{15^+,k}(t)$ is the number of people in location k in year t who are at least 15 years old, $p_{15^-,k}(t)$ is the number of people in location k in year t who are less than 15 years old, and $F_k(t)$ is the overall average household size in location k in year t.

Two aspects of these specifications deserve attention. First, it appears that equations (3.25) and (3.26) by themselves should be sufficient to determine the composition of households in the rural and urban areas by income level. One difficulty with them is that the predicted numbers of adults and children when summed are inconsistent with the aggregate family size. This situation requires the adjustment made in equation (3.27). Unfortunately, the equations for that adjustment do not appear in the monograph. If the adjustment is like the others made in the model, it would be a proportional increase or decrease in all the relevant figures.

This leads to the second point. Adjustments consistent with equation (3.27) can be demographically inconsistent. A preferable way of proceeding would be to make some kind of adjustment that maintains the following two basic identities:

$$\sum_{d=1}^{10} A_{k,d}(t) = p_{15^+,k}(t) \tag{3.28}$$

and

$$\sum_{d=1}^{10} C_{k,d}(t) = p_{15^-,k}(t) \tag{3.29}$$

Adjustments performed to ensure that equation (3.27) holds do not necessarily ensure that the two identities above are met. This results in the possibility that the number of adults and children by income level do not necessarily aggregate to the number of adults and children in the relevant population group. This problem affects both the segment of the model dealing with the distribution of income and the portion dealing with savings and consumption. Policy makers wishing to use the Bachue framework should substitute a specification here that ensures that the demographic aggregation is correct.

Education

The Bachue model distinguishes three levels of education, less than primary, primary completed and less than secondary, and secondary completed or more. The major assumptions in the specification are that all children are enrolled in primary school and that their progression through the educational system is determined by a set of governmentally controlled completion rates. To the extent that completion rates are truly exogenous and under the control of the government, this specification is sufficient for modeling purposes. It should be noted in passing, however, that formal schooling is the only route to the development of skills in the labor market. If this is not roughly true in a country of interest to the policy maker, he could expand the specification given in the model.

Fertility

The fertility variable endogenously explained in the Bachue model is the gross reproduction rate. The authors of Bachue reject microlevel fertility equations and use instead an equation estimated on country-wide data. This choice is likely to be a wise one. Microeconomic and microdemographic specifications are unlikely to yield an equation that can predict a demographic transition, but country-wide data may be useful in this regard. The equation in the model has the following form:

$$\begin{aligned} GRR_k(t) = {} & b_k - 0.0064 \cdot R_k(t-1) + 0.0106 \cdot I_k(t-1) \\ & - 0.0446 \cdot e_k^0(t-1) + 0.0059 \cdot L_A(t-1) \end{aligned} \qquad (3.30)$$

where b_k is 5.14 in urban areas and 5.19 in rural areas, $R_k(t-1)$ is the female labor force participation rate in region k in year $t-1$, $I_k(t-1)$ is the percent illiterate in region k in period $t-1$, $e_k^0(t-1)$ is the life expectancy at birth in region k in year $t-1$, and $L_A(t-1)$ is the proportion of employment in agricultural activities (presumably in the country as a whole) in year $t-1$. It should be noted that this specification assumes that the government has no direct role in lowering fertility through programs of dissemination of contraceptive information or devices.

The demographic segment of the model, however, requires a set of age-specific marital fertility rates. Unfortunately, in moving from the gross reproduction rate to these rates, an error of disaggregation is made. The age-specific marital fertility rates are derived as follows

$$MF(a, t) = k_1(a) + k_2(a) \cdot TFR(t) + k_3(a) \cdot M(a, t) \qquad (3.31)$$

where $MF(a, t)$ is marital fertility at age a in period t (in the model the dependent and independent variables are also specific for urban and rural location), $TFR(t)$ is the total fertility rate in period t (it equals the gross reproduction rate multiplied by a known constant), $M(a, t)$ is the proportion of women of age a in period t who are married, and the $k_i(a)$ are age-specific constants. The difficulty with this formulation is that the computed marital fertility rates, when used with the proportions married estimated in the model (see the following subsection) do not necessarily aggregate correctly to the total fertility rate. To make this point more precisely, let us write the relationship between marital fertility rates, marriage rates, and the total fertility rate:

$$\sum_a MF(a,t) \cdot M(a,t) = TFR(t) \tag{3.32}$$

where the summation is taken across the reproductive ages.

Multiplying equation (3.31) by $M(a, t)$ yields

$$\begin{aligned} MF(a,t) \cdot M(a,t) &= k_1(a) \cdot M(a,t) + k_2(a) \cdot TFR(t) \cdot M(a,t) \\ &\quad + k_3(a) \cdot M(a,t)^2 \end{aligned} \tag{3.33}$$

Summing over the reproductive years, rearranging terms, and utilizing equation (3.32) produces the equation

$$TFR(t) = \frac{\sum_a [k_1(a) \cdot M(a,t) + k_3(a) \cdot M(a,t)^2]}{\sum_a [k_2(a) \cdot M(a,t)]} \tag{3.34}$$

The meaning of equation (3.34) is clear enough. Given the marriage rates and the age-specific constants in equation (3.31), the total fertility rate and therefore the gross reproduction rate are determined. To put the matter somewhat differently, under those conditions equation (3.30) is redundant and in general contradictory. Of course, we can consider the gross reproduction rate as calculated in equation (3.30) to be correct. In that case, one of the marriage rates must be determined by equation (3.34). Unfortunately, as is described in the following subsection, all the marriage rates are computed independently. Clearly, equation (3.31) introduces an inconsistency into the model – the age-specific marital fertility rates, marriage rates, and gross reproduction rate are overdetermined. This problem should be alleviated before the impacts on fertility of various policy changes are analyzed.

Marriage Rates

Marriage rates play two roles in the Bachue model, one relating to the determination of female labor force participation, the other relating to fertility. The mean ages at marriage in the rural and urban sectors are determined from linear equations where the dependent variables are the change from 1965 to the current year in the proportion of women with primary education not completed, the change from 1965 to the current year in the proportion of women aged 15–29 with secondary education, and the change from 1965 to the current year in the proportion of women 15–29 in the labor force.

Given the mean age at first marriage, the authors claim to obtain the age-specific proportions married from the standard nuptiality rate table in Coale (1971).[20] Technically speaking, however, that work cannot be used to determine age-specific proportions married but rather age-specific proportions of women *ever* married. This difference is not of much importance where life expectancies are relatively high and where divorce rates are relatively low, but it may be of some importance where these conditions are not met. Even as an approximation, the Coale nuptiality rate formulation is the best possible one to use in this context.

Mortality

Life expectancy at age zero in the rural and urban areas is derived from a linear function of three variables: (a) the inverse of per capita gross domestic product, (b) the inverse of the square of per capita gross domestic product, and (c) the Gini coefficient of income inequality. A separate life expectancy is computed for the urban and for the rural areas. Given the life expectancy at age zero, age-specific mortality rates are determined by using the Coale and Demeny (1966)[21] model West life tables. Although the West tables are probably the most accurate of the Coale and Demeny model life tables, they are not particularly well suited to the Philippine case. The underlying data for those tables come predominantly from high-income, low-fertility countries. The experience of low-income, high-fertility countries is probably captured more appropriately by the model South life tables.

Migration

Bachue is the only model among those reviewed here that deals with gross as well as net migration flows. The gross flow of migration from rural areas to urban areas is decomposed by age, sex, and education, as is the return flow from urban areas to rural areas. The gross rate of migration (specific for age, education, and sex) is given by the product of three terms. The first term depends upon the proportion of women married at the given age, average educational level, distribution of education, and age. The second term depends upon the relative wage rates in the urban and rural areas and upon the coefficients of variation of income in the two regions. The third term varies with the locational distribution of the population. For rural-to-urban migration, it assumes that the propensity to migrate increases until 50 percent of the population is urban and decreases thereafter. For urban-to-rural migration, it is assumed that the propensity to migrate decreases as the proportion of the population in urban areas grows.

3.5 CONCLUSIONS

The Bachue model is, in its present form, of little use to agricultural policy planners. This is the case for two basic reasons. First, the model is not designed to focus on agriculture. It is, therefore, not sufficiently articulated with regard to agriculture to provide interesting policy options for study. For example, inputs into agricultural production have no effect on the level of agricultural output. Thus, the whole set of questions concerning the relations between agricultural outputs and inputs cannot be addressed in the model. Certainly, it is possible to change the rate of labor productivity growth in the rural area, but without some understanding of what is generating this growth, the formulation is not very useful for agricultural planning.

The second reason that the Bachue model may not be very instructive in its present version is that it contains a number of difficulties, some of which are purely technical. For example, the wrong prices are used in the consumption equations, the translation between value-added in constant prices and value-added in current prices is incorrectly made, and the age-specific marital fertility rates and the proportions married at given ages are inconsistent with the total fertility rate in the model. These technical problems need to be remedied before the results of the model can be taken seriously. There is, however, a more basic problem that needs to be considered. Bachue is inherently a demand-dominated model. The supply side of the model – the relationship between output levels and input levels – is assumed to have almost no role in the growth process. The thrust of the model is to explain not the pattern and speed of development under various assumed policies, but to determine the consequences of those policies for the distribution of income. It is legitimate, of course, to ask about the effects of various policies on the distribution of income in the rural sector. But in a model where there are no effects of those policies on growth and, practically speaking, no effects on value-added shares, it is not obvious whether those questions can realistically be answered.

The Bachue model should be applauded for its serious consideration of questions concerning income distribution, but it must be remembered that this focus has been achieved at the expense of other important considerations. Even when the technical problems are resolved, a planner may well have second thoughts about using the model. Merging the income distribution considerations in the Bachue model with the supply elements of other models may be a very useful tack for policy makers interested in models of this kind.

4 THE TEMPO II MODEL

4.1 THE PRODUCTION RELATIONS

Tempo II is among the simpler models considered in this review. It recognizes only two sectors of the economy, a subsistence sector and a modern sector. The production function for the subsistence sector (in essence the agricultural sector) is

$$GPS(t) = k_1 \cdot PS(t-1)^{k_2} \qquad (4.1)$$

where $GPS(t)$ is the gross product of the subsistence sector in year t, $PS(t-1)$ is the size of the population (not labor force) in the subsistence sector in period $t-1$, and k_1 and k_2 are constants. The authors of the Tempo II model suggest that k_2 should be less than unity in order to ensure that labor in the subsistence sector always faces diminishing returns.

Several comments on this specification are in order before we move on to a discussion of the production function in the modern sector. First, the agricultural production structure is extremely simplified. Land and agricultural capital are assumed to have no relation to agricultural output. Further, even in the simple two-sector economy of Tempo II there are no intersectoral purchases. In other words, fertilizer or electricity purchased from the modern sector are not inputs into agricultural production. Such a view of agriculture may be based on a perception that agriculture in some less developed countries is carried out with little more than land and labor. Although this may or may not be true today for any given country, it should not be forgotten that economic–demographic simulation models are designed to run for 20 to 30 years into the future. In this perspective, omitting all agricultural inputs except labor from this agricultural production function is not a very convincing assumption. Just as bad, however, is the assumption that there will be no technological change in agriculture over the course of the next two or three decades.

The single factor of production in the agricultural production function is the *population* in the subsistence sector. No attempt is made in Tempo II to define an agricultural labor force or to use information on the age structure of the agricultural population to adjust the population to a number of full-time equivalent workers. The Tempo II approach requires little in the way of data, but also seems to offer little in return. One curiosity of this approach is that agricultural output in period t is assumed to be a function of the agricultural population in period $t-1$. This eliminates the problem that could arise if this year's agricultural output is determined simultaneously with this year's migration flow. But this solution has a cost in terms of the realism of the model.

The production function for the modern sector in Tempo II is written as

$$GPM(t) = Z \cdot (1+q)^t \cdot K^u(t-1) \cdot NE^v(t-1) \cdot NU^w(t-1) \qquad (4.2)$$

where $GPM(t)$ is the output of the modern sector in year t, $K(t-1)$ is the capital stock in the modern sector in period $t-1$, $NE(t-1)$ is the number of employed educated workers in period $t-1$, $NU(t-1)$ is the number of employed uneducated workers in period $t-1$, q is a constant reflecting the rate of technological progress, and Z, u, v, and w are also constants. The authors of Tempo II say nothing about restricting the sum of u, v, and w.

It is somewhat curious that the output level in period t depends on input levels in period $t-1$. This formulation makes the modern-sector output in period t completely independent of any economic phenomena in period t. Such a specification may be useful for certain purposes, but it certainly detracts from the realism of the model. It should also be noted here that the modern sector does not utilize anything from the subsistence sector in its own production. For example, the modern sector is not allowed to process food, nor are businessmen in the modern sector allowed to purchase items produced in the subsistence sector for export.

The three inputs used in modern sector production in period t are all determined in period $t-1$. The level of the capital stock in period $t-1$ is easily computed since it is assumed to include all the capital in the entire economy. Tempo II does not allow for such items as roads, fences, or buildings in the agricultural area. The number of employed educated workers in period $t-1$ is calculated as a fixed exogenous proportion of the number of educated people in the labor force in period $t-1$. In other words, the rate of unemployment among educated workers is assumed to be constant in Tempo II.

The determination of the number of uneducated workers employed in period $t-1$ is somewhat more complex. Basically, their number is determined from the following equation:

$$NU(t-1) = \kappa_1 - \kappa_2 \cdot \left[\frac{LFU(t-2)}{K(t-2)}\right] \cdot LFU(t-1) \qquad (4.3)$$

where $NU(t-1)$ is the number of uneducated laborers employed in the modern sector in year $t-1$ (but somehow producing output in year t), $LFU(t-1)$ is the size of the labor force of uneducated workers in the urban area in period $t-1$, and κ_1 and κ_2 are two positive constants. Equation (4.3), unfortunately, is quite implausible. The problem with that specification can be demonstrated in a simple example. Consider for the moment two years t and $t+20$ and allow the urban unskilled labor force and the capital stock to have grown at the same rate over that period, or, in other words, let $LFU(t)/K(t)$ remain constant over time. In this case equation (4.3) may be rewritten

$$NU(t) = \kappa_1 - \kappa_2^* \cdot LFU(t) \qquad (4.4)$$

where

$$\kappa_2^* = \kappa_2 \cdot \left[\frac{LFU(t-1)}{K(t-1)}\right] \qquad (4.5)$$

Now in year t suppose $LFU(t) = 100$ and $NU(t) = 90$. One way to obtain this result is to set $\kappa_1 = 100$ and $\kappa_2^* = 0.1$. If $LFU(t+20) = 200$, we would have the astounding implication that $NU(t+20) = 80$. In other words, while the labor force doubled, employment shrank by about 11 percent. Indeed, this negative relation between employment and the labor force is evident from equation (4.3). What, if any, sense the equation makes eludes this author.

Tempo II incorporates one innovative feature with regard to the determination of the number of uneducated workers employed in the modern sector. It is an adjustment for the increased "quality" of uneducated labor that comes about over time with development because of increased nutritional levels and decreased morbidity. The equation which incorporates this adjustment is

$$LFUA(t) = LFU(t) \cdot \left\{\left[\frac{GPM(t-1)}{PM(t-1)}\right] \Big/ \left[\frac{GPM(0)}{PM(0)}\right]\right\}^h \qquad (4.6)$$

where $LFUA(t)$ is the adjusted labor force size in period t (i.e., the number of equivalent workers given the health and nutritional standards of period 0), $PM(t-1)$ is the total population in the urban sector in period $t-1$, and h is a constant bounded by zero and unity.

The authors of Tempo II suggest that the value of $LFUA(t)$ in equation (4.6) can be used in place of $LFU(t)$ in equation (4.3), but because of the problem with equation (4.3), this novel feature of Tempo II may only serve to compound the poor specification.

4.2 THE DISTRIBUTION OF INCOME

Unlike the other models reviewed here, Tempo II makes no distinction between income generated in the urban area and income generated in the rural area, nor is any distinction drawn between labor and nonlabor income. Tempo II recognizes only a single aggregate form of income. Disposable income in period t is computed according to the following equation:

$$DI(t) = GP(t) - TAX(t) + TRFP(t) \qquad (4.7)$$

where $DI(t)$ is disposable income in period t, $GP(t)$ is gross national product in period t [$= GPM(t) + GPS(t)$], $TAX(t)$ is the sum of all taxes in period t, and $TRFP(t)$ is the sum of all transfer payments in period t. The Tempo II definition of disposable income thus includes all business income and the depreciation on the entire capital stock.

Because Tempo II virtually ignores the distribution of income, it is not useful for analyzing policies where changes in the distribution of income are likely to be sizable. For example, it may be argued that increases in population growth tend to depress wage rates and increase the shares of profits and rents in national income. If savings rates out of profits and rents were higher than out of wage income, more rapid population growth could cause the aggregate savings rate to rise. None of this story can be captured in Tempo II. This failure to deal with the income distribution is a significant deficiency of Tempo II.

4.3 SAVINGS

The aggregation of all incomes in the Tempo II model limits the sophistication of the savings process. In Tempo II the ratio of savings to disposable income is expressed in the following relationship

$$\frac{S(t)}{DI(t)} = a_1 - a_2 DI(t)^{a_3 - 1} P(t)^{1-a_3} \qquad 0 \leqslant a_3 \leqslant 1 \qquad (4.8)$$

where $S(t)$ is aggregate savings in year t, $DI(t)$ is disposable income in year t, $P(t)$ is the size of the population in year t, and a_1, a_2, and a_3 are constants. If a_3 lies in the interior of the unit interval, this equation implies that the savings rate is positively related to disposable income per capita.

The difficulty with this specification is not so much with what it maintains as with what it fails to consider. For example, changes in the distribution of income toward either large firms in urban areas or large farmers in the rural area are assumed to have no effect on savings. Nor does the rate of interest or the rate of inflation – a variable uniquely available in Tempo II – have even the slightest impact on the savings rate. Likewise, the age structure of the population, its rural–urban composition, and its educational distribution all have no impact on savings. In Tempo II, the trend in the savings rate is simply determined by the trend in per capita disposable income.

4.4 THE DETERMINANTS OF FINAL DEMAND

In Tempo II, there is a single equation for calculating the aggregate level of consumption expenditures that may be written

$$C(t) = (1 - a_1)DI(t) + a_2 DI(t)^{a_3} \cdot P(t)^{1-a_3} \qquad 0 \leqslant a_3 \leqslant 1 \qquad (4.9)$$

where $C(t)$ is aggregate consumption in year t and where the constants $a_1, a_2,$

and a_3 are the same as in equation (4.8). If a_3 is not unity, consumption increases less than proportionally with disposable income.

It would perhaps have been redundant at this point to comment on the lack of relative prices in Tempo II, except that this attribute of the model, together with its supply-constrained character, yields a rather unfortunate result here. In Tempo II consumption is not disaggregated even into the demand for the two outputs considered in the model. The reason that consumption of agricultural goods is not differentiated from the consumption of goods and services produced in the modern sector is simple enough. Without any relative price in the model, there is no method of ensuring that the demand for either sector's output in any period will be equal to the exogenous quantity produced in that period.[22]

From the perspective of a policy maker the level of aggregation of consumption in Tempo II is likely to cause significant difficulties. Tempo II does not allow the analysis of any policy that involves encouragements or discouragements to output growth from the demand side. For example, Tempo II is incapable of analyzing the direct or indirect effects of a subsidy to agricultural production, of a tax on modern-sector outputs, or even of a tariff on competitive imports. This is certainly one area in Tempo II which should be expanded significantly before the model is used for serious work.

Another shortcoming of the Tempo II model is that it contains no independent specification of the demand for investment. Investment is determined, in Tempo II, from the accounting relationship

$$PINV(t) = S(t) - BOR(t) \tag{4.10}$$

where $PINV(t)$ is private investment at time t, $S(t)$ is savings at time t, and $BOR(t)$ is government borrowing in period t. If the government's deficit were entirely met by domestic borrowing, then equation (4.10) would guarantee that aggregate demand was equal to aggregate supply.

In an economic-demographic simulation model, however, determining private investment from an accounting relationship is inappropriate because this procedure leads to the omission from the model of all factors that influence the process of growth and development by affecting the profitability of investment. Tempo II is thus incapable of analyzing any policy that works through the stimulation of investment. It would be a considerable improvement in Tempo II if the elementary distinction between the determinants of *ex ante* and *ex post* investment was made.

The allocation of investment between sectors in Tempo II has been reduced to a trivial problem by assuming that there is only one capital stock – the capital stock of the modern sector. Tempo II, then, cannot be used to analyze policies that have the effect of redirecting private investment among sectors of the economy.

Government expenditures in Tempo II are divided into eight categories: education, family planning, general transfer payments, health, social overhead capital, direct government investment, defense, and general government. Expenditures in each of these areas except education and family planning are exogenous policy variables and may be changed over time.

Expenditures on education and family-planning services are computed within a goal-oriented framework. There, instead of specifying the amount of money to be spent on a given social program, the policy maker sets the target levels of educational attainment and fertility reduction he wishes to achieve and the simulation model determines both the cost of achieving each goal and the impacts on the development process of reaching those ends. The costs and benefits of pursuing various policies can be more easily seen in this framework than where government expenditures are treated as being purely exogenous. This aspect of the Tempo II model is one that could profitably be incorporated into the next generation of economic–demographic simulation models.

Government revenue is determined in Tempo II through the use of the equation

$$TAX(t) = \tau \cdot GP(t) \tag{4.11}$$

where $TAX(t)$ is government revenue from taxation in period t, $GP(t)$ is gross national product in period t, and τ is the tax rate. The deficit in the government budget is simply the difference between its revenues from internal taxation and government expenditures. This deficit is financed in an intriguing manner in the world of Tempo II. It is assumed that small deficits are entirely covered by borrowing from domestic savings. The size of the budget deficit that can be financed in this fashion is limited to some fixed fraction of total domestic savings. If the deficit exceeds the limit, the excess is financed in essence by the creation of new money, thus causing inflationary pressure. The treatment of general inflation in Tempo II is discussed in section 4.5 below.

Neither exports, imports, nor capital flows are incorporated into the Tempo II framework.

4.5 GENERAL EQUILIBRIUM ASPECTS

Tempo II is a supply-dominated model. Output in the current period is derived from the quantities of the factors of production determined in the previous period. Since consumption and investment are treated as national aggregates in Tempo II, it would seem to be a simple matter at first to guarantee that *ex ante* aggregate demand equaled *ex ante* aggregate supply. This is especially true since the *ex post* equilibrium condition that aggregate savings is equal to aggregate investment is invoked as an *ex ante* relationship determining the amount of aggregate investment. Tempo II, however, incorporates two features that allow *ex ante* aggregate demand to differ from *ex ante* aggregate supply. First, the government can cover a portion of its deficit by printing money. In the absence of external aid this causes the aggregate demand for goods and services in any given year to exceed the quantity of goods and services produced in that year. The second reason why *ex ante* aggregate supply and demand may deviate from one another involves the existence of long-term external aid. In Tempo II, long-term external aid supplements the domestic supply of goods and services. If the government does not finance its deficit by adding to the supply of money, such aid causes a tendency for aggregate supply to exceed aggregate demand.

Of course, *ex post* aggregate demand must equal *ex post* aggregate supply, and for this Tempo II includes a mechanism that guarantees the *ex post* equality even when the *ex ante* equality does not obtain. Let us explore this mechanism for a moment. Suppose that the ratio of *ex ante* aggregate supply (including long-term external aid) to *ex ante* aggregate demand in year t is given by $R(t)$. Since Tempo II is a supply-dominated model, the discordance between *ex ante* aggregate demand and aggregate supply is eliminated by multiplying all the elements on the income side of the national accounts (for example, disposable income, consumption, investment, and government spending) by $R(t)$. As the authors of Tempo II suggest, this simple strategem can be made more sophisticated by positing that the components of final demand are affected to varying degrees in the course of aggregate demand–supply adjustment.

The economic logic of modifying *ex ante* aggregate demand so that it comes into equality with aggregate supply is not treated in detail in Tempo II, but there is a suggestion of the mechanism by which at least part of the adjustment takes place. In Tempo II, the rate of inflation between periods $t - 1$ and t is described by the equation

$$INFL(t-1, t) = R(t)^{\alpha} - 1 \qquad \alpha > 0 \tag{4.12}$$

where $INFL(t - 1, t)$ is the rate of inflation between periods $t - 1$ and t, and α is a constant. If *ex ante* aggregate demand exceeds aggregate supply $[R(t) > 1]$, then inflation occurs, and if *ex ante* aggregate demand falls short of aggregate supply $[R(t) < 1]$, then deflation follows. It may easily be imagined that changes in the rate of price inflation could play a role in the adjustment of aggregate demand to aggregate supply, but the precise nature of this role is left unspecified in Tempo II. Perhaps in further work on this model, the link between inflationary pressures and changes in real quantities demanded can be better articulated.

It should be noted before we leave this topic that while the treatment of inflation is hardly complete (for example, the effect of inflation upon the savings rate is not considered), it is at least a first recognition of a basic fact of life in many developing countries. It is interesting to observe in this context that in Tempo II long-term external aid tends to have a deflationary effect on the economy because it adds to aggregate supply without affecting government policies. Were some link made between long-term aid and monetary expansion, this deflationary effect could disappear.

4.6 THE DEMOGRAPHICS

The Demographic Accounting

The demographic accounting in Tempo II is done on a cohort basis. The framework distinguishes people by sex, by location, and by single years of age from age 0 to the ages 65 and above. It presumably should also classify people according to their educational attainment, but no mention is made of this. Very little of the age detail is used in the economic portion of the model.

Labor Force Participation Rates

In Tempo II no attempt is made to define the agricultural labor force. In essence, it is considered to be the entire agricultural population. Labor force participation rates of educated and uneducated workers in the urban sector are determined in a comparably simple fashion. It is assumed in Tempo II that age-, sex-, and education-specific labor force participation rates are fixed constants invariant both to policy manipulation and to economic and demographic developments. The educated and uneducated urban labor forces are determined by applying these exogenous labor force participation rates to the numbers of people in the relevant age, sex, and education categories.

The treatment of labor force participation rates in Tempo II is a particularly simple one. Before a policy maker can be expected to believe 20- or 30-year simulations based on the assumption that labor force participation rates by age, sex, and education will not change over that span, he deserves some relevant empirical evidence on this point. Without such a demonstration, he may properly remain skeptical of this portion of the model.

Education

Education in Tempo II, like family planning, is treated as a special service in that the government is assumed to have target (age- and sex-specific) enrollment rates for primary, secondary, tertiary, and professional education. The only question that arises, then, is how much all this education is going to cost. The problem of providing the education does not arise in the model. To simplify the story slightly, the cost of education (in base-year prices) may be written as

$$TCE(t) = \sum_{i=a_0}^{a_1} [EN(i,t) \cdot P(i,t) \cdot ce(i)] \tag{4.13}$$

where $TCE(t)$ is the total cost of education in year t, a_0, and a_1 are the initial and terminal ages of public education, $EN(i, t)$ is the exogenous enrollment rate for people of age i in year t, $P(i, t)$ is the total number of people of age i in year t, and $ce(i)$ is the cost in base year prices of educating an i-year-old person.[23] The real costs of a year of education at each age level are assumed to remain constant.

There are several puzzling aspects of the education specification in the context of the full model. The most immediate question concerns the constancy over time of the real cost of providing a year of schooling at each level. The educational system uses skilled manpower intensively, and one would expect that the real cost of a year of schooling would be affected by the real earnings of educated workers. As a country developed, one would expect both an increase in the real earnings of educated workers and an increase in the real cost of education. The assumption in Tempo II that the real cost of education remains fixed over time is liable to suggest to policy makers that development strategies involving increasing human capital are quite a bit less costly than they are likely to be in reality.

In Tempo II, there are two types of labor in the modern sector, educated and uneducated labor. Yet the schooling system potentially produces people with quite a variety of educational backgrounds. The relation between this array of schooling levels and the bipartite distinction between educated and uneducated labor is unclear in Tempo II. Surely one can easily imagine classifying anyone with *n* years of schooling or more as an educated worker and anyone with fewer years of schooling as an uneducated worker, but any such classification may produce highly misleading results in the simulations. For example, if after 20 years of sustained effort most of the workers could be classified as educated workers, further expenditures on education may appear to have a spuriously low return because few additional people are being moved from the category "uneducated" to the category "educated." More disaggregation by educational level would be useful here.

Fertility and Family Planning

Fertility is treated in very simple fashion in Tempo II. The model uses sets of age-specific fertility rates for the urban and rural areas and derives the number of births in any year by applying these rates to the relevant numbers of females by age and summing across the reproductive age span. This approach is a good one thus far, but the most important issue is the determinants of the age-specific fertility rates. Here Tempo II is extremely weak. These fertility rates are treated as if they were influenced by only one variable in the model, family-planning expenditures. Education, income, mortality rates, and health conditions have no impact on fertility in the world of Tempo II.

Even the specification of the impact of the family-planning program is very limited in Tempo II. The family-planning program in Tempo II is assumed to cover only females in the urban sector. Thus, over the entire 20- to 30-year simulation period, the government is prevented from providing any family planning services in rural areas. This assumption is dubious for many developing countries. Since family-planning expenditures are the only determinant of age-specific fertility rates in the Tempo II framework and there are no such expenditures permitted in rural areas, rural fertility rates are completely exogenous in Tempo II. For a model whose use is to provide information about the relationships between economic and demographic variables, this specification is egregious.

In Tempo II, the proportion of fertile urban women using contraception affects the number of births according to the equation

$$\frac{BU(t)}{BU^*(t)} = 1 - PU(t-1) \tag{4.14}$$

where $BU(t)$ is the actual number of births in the urban area in time period t, $BU^*(t)$ is the hypothetical number of births that would have occurred in the urban area in time period t had no contraception been employed, and $PU(t-1)$ is the proportion of fertile urban women who were using contraception in period $t-1$.[24] Since the proportion of users is an exogenous policy variable, the government has the power to reduce urban fertility to any level it chooses. The only constraining factor is the cost of this fertility reduction.

Given the mandated proportion of urban women between the ages of 15 and 49 using contraception and the number of these women, the cost of the fertility reduction is determined by the average cost per user. It is assumed in Tempo II that all such costs are borne by the government. The annual real cost to the government of an urban woman using contraception is assumed to be constant as long as the rate of use is below some critical value. When the rate exceeds the critical value the annual real cost to the government per user is assumed to increase linearly with the rate of contraceptive use. Clearly this is an *ad hoc* formulation. A policy maker should carefully consider whether such a framework is appropriate for his country over a 20- to 30-year horizon.

Mortality Rates

All mortality rates in Tempo II are assumed to be exogenous. Neither government public health projects nor rising levels of income and education are allowed to have any effect on mortality rates.

4.7 DYNAMIC CONSIDERATIONS

Economic growth and development occurs in the Tempo II model because of technological progress in the modern (urban) sector, labor force growth, the growth of the stock of educated manpower, the growth of the capital stock, and the reallocation of unskilled labor from the rural sector to the urban sector. Most of these processes are treated quite simply in Tempo II. Technological change in the urban modern sector is both Hicks- and Harrod-neutral and occurs at a constant exogenous rate. The stock of (urban) capital grows through the annual addition of net investment. There is no problem of allocating investment funds between competing uses because only a single aggregated capital stock appears in the model. The stock of educated manpower grows at a rate determined by the government, and, since the education of rural residents does not affect agricultural output, the question of intersectoral educational strategies does not arise.

The migration specification in the Tempo II model is also reasonably simple. It is assumed that the annual flow of migration can be determined from the following equation:

$$M(t) = \alpha[r(t-1)]^{\beta} \cdot PS(t) \tag{4.15}$$

where $M(t)$ is the net migration from rural to urban areas in period t, $r(t-1)$ is the ratio of the income of employed unskilled workers in the urban area in period $t-1$ to the average output of all members of the agricultural population in period $t-1$, $PS(t)$ is the number of people in rural areas in period t, and α and β are constants. The ratio $r(t-1)$ is defined as

$$r(t-1) = \left[\frac{w \cdot ZM(t-1)}{PMU(t-1)}\right] \Big/ \left[\frac{ZS(t-1)}{PS(t-1)}\right] \tag{4.16}$$

where $ZM(t-1)$ is the output of the modern (urban) sector in period $t-1$, $ZS(t-1)$ is the output of the subsistence (rural) sector in period $t-1$, $PMU(t-1)$ is the number of unskilled workers in the modern sector in period $t-1$, $PS(t-1)$ is the number of people in the subsistence (rural) sector at time $t-1$, and w is share of the value of output paid to unskilled workers in the modern sector.[25]

There are several debatable features of this migration specification that need to be brought to the attention of its potential users. Let us start with the simplest problem and progress toward more subtle ones. In equation (4.15), the migration stream and the rural population base from which it derives have the same date. The question that must be answered here is whether the rural population in period t includes or excludes the migrants in period t. The answer in turn has implications for other equations in the model.

Several more substantive issues arise concerning the rate of rural-urban migration. First, the rate of rural outmigration is assumed to be independent of the age and sex structure of the rural population. Thus, a rural population with a large proportion of young adults in their late teens and early twenties will, in Tempo II, have the same migration rate as a population composed dominantly of elderly people. Such a formulation is not terribly realistic. Further, the rate of migration in period t is assumed to depend only on conditions in period $t-1$. Whether this is an appropriate simplification may depend on the particular application.

Another problem with the migration rate formulation is that it does not recognize the existence of migration costs. Let us assume for the moment that $r(t-1)$ correctly measures the relevant incomes of potential migrants. When $r(t-1) = 1$, there is no economic incentive for migration to continue, yet the rate of rural outmigration will be greater than zero. Indeed, migration to urban areas will continue even when rural incomes exceed urban (unskilled) incomes by a considerable margin. Migration will stop only when the average income of unskilled workers in the urban areas goes to zero. Given a Cobb-Douglas production function for the output of the urban modern sector, zero average income of unskilled workers can occur only when output is itself zero. Thus, in the Tempo II model, the existence of nonzero output in the urban area guarantees migration from rural to urban areas even if wages are higher in the countryside than in the city. A more plausible specification such as that found in the KWC model forces migration to a halt when the difference between urban income and rural income falls below some critical value.

In addition to its failure to recognize costs of migration, the Tempo II model also fails to make a distinction between output measured in physical terms and the value of output. It is natural to think that, for potential migrants, one attraction of urban areas is the higher level of income there. The ratio $r(t-1)$ in equation (4.15) is supposed to capture this effect, but it does not if the relative prices of rural and urban sector outputs change with development. In equation (4.15) $r(t-1)$ is the ratio of two numbers of *physical units*, not the ratios of two income levels. If the relative price of the outputs remain unchanged, the output ratio will serve as an acceptable proxy, but if the terms of trade change over time, $r(t-1)$ will no longer serve as a proxy for the proper income ratio and migration will be poorly predicted.

4.8 POLICY QUESTIONS

The Tempo II model is not suited for the analysis of any questions concerning the agricultural sector. The government cannot encourage technological progress in agriculture because it is assumed that there is no technological progress in agriculture. The government cannot improve the productivity of agricultural labor through education because it is assumed that education has no influence on the productivity of rural laborers. The government cannot increase agricultural output through the provision of social overhead capital in the rural area because the Tempo II model does not include an agricultural capital stock. The government cannot directly influence the rate of population growth in the rural areas because the model assumes that all family planning expenditures are made in the urban areas.

What questions then can be addressed meaningfully in the Tempo II framework? It is sensible to ask only about certain aspects of family-planning programs and educational policy – but even in these limited areas the answers are not very informative. For example, one need not actually perform the simulations to observe that, in the context of Tempo II, increases in expenditures on family planning almost automatically bring about an increase in per capita income. To see this, consider an economy with an average per capita income of \$500 where the elasticity of the output of the modern sector with respect to its capital stock is 0.25. Further, let us consider the effect of an expenditure of an additional $\$X$ on the family-planning program in year t where $\$X$ is the amount required to avert one birth. In year $t + 1$ the population is one person lower than it otherwise would have been (for simplicity, mortality is ignored here) and the capital stock is $\$X$ lower than it otherwise would have been. Output in the modern sector, however, is approximately only $(0.25) \cdot (\$X)$ less than it would have been. If $(0.25) \cdot (\$X)$ is greater than the per capita income of \$500, the expenditure on the family-planning program would have caused a diminution in real per capita income, and if, on the contrary, $(0.25) \cdot (\$X)$ is less than \$500, per capita income would have increased. The crucial point is that family planning expenditures immediately increase per capita income if the cost of averting *one* birth is less than \$2,000 or less than four times the average per capita income in the country. Since any family-planning program is likely to require less than four times the average per capita income to avert a single birth, the short-run effect of family planning expenditures is clearly a foregone conclusion.

The longer-term implications of reducing fertility all work in the same direction. A smaller population is associated with a higher savings rate, faster rate of growth of the urban capital stock, and, therefore, higher urban wage rates for unskilled workers. This causes migration from rural areas to urban areas to increase, and, since the marginal product of labor is higher in the urban areas than in the rural areas, it causes, in turn, an increase in per capita income. Thus, the specification of Tempo II essentially builds in the conclusion that increases in expenditures on family-planning programs cause increases in per capita income.

In brief, the framework of Tempo II is not sufficiently articulated to provide the policy maker with much valuable information about the direct or indirect effects of policy changes.

5 THE SIMON MODEL

5.1 PRODUCTION RELATIONS

In the Simon model, there are two types of goods produced, industrial-sector output and agricultural-sector output. Industrial output is specified as resulting from the Cobb–Douglas production process

$$Q_I(t) = A_I(t) \cdot K_I^{0.4}(t) \cdot M_I^{0.6}(t) \cdot J(t) \qquad (5.1)$$

where $Q_I(t)$ is industrial output in time period t, $A_I(t)$ is the value of the industrial "technology" index in period t, $K_I(t)$ is the industrial capital stock in period t, $M_I(t)$ is the number of man-hours of labor spent in the industrial sector in period t, $J(t)$ is an index of the quantity of social overhead capital in the country as a whole in period t.[26] The agricultural production function is also Cobb–Douglas. It is expressed as

$$Q_F(t) = A_F(t) \cdot K_F^{0.5}(t) \cdot M_F^{0.5}(t) \cdot J(t) \qquad (5.2)$$

where the variables are defined analogously to those in the industrial production function, with the exception that $K_F(t)$ includes land.

The Simon model, then, allows for neutral technological change in both the agricultural and the industrial sector and formally treats the role of social overhead capital in production. The motivation behind this specification is to be applauded. For all the discussion in the literature about the role of the government in providing social overhead capital, the Simon model is the only one of those considered here that treats this form of capital explicitly. The details of the incorporation of social overhead capital into the model, however, leave something to be desired. First, the social overhead capital variable $J(t)$ enters both production functions with an exponent of unity. In other words, it is possible to double or quadruple output in both sectors of the economy by doubling or quadrupling social overhead capital without *any* increase in the utilization of labor or the services of the private capital stock. Whether social overhead capital has such a potent effect on output remains to be demonstrated.

An economist's presumption would be that social overhead capital, like any other input, would eventually encounter diminishing returns to scale. It should also be noted in passing that the stock of social overhead capital is not disaggregated by sector. Thus the building of a rural road will not only increase rural output, but will directly increase industrial output as well.

This process by which social overhead capital is assumed to grow is also rather puzzling. Simon writes that

$$\frac{J(t+1)-J(t)}{J(t)} = 0.20 \cdot \left[\frac{L(t)-L(t-1)}{L(t-1)}\right] \tag{5.3}$$

where $L(t)$ is the labor force in the entire country in period t. The stock of social overhead capital, according to this formulation, automatically grows whenever the labor force grows. No difficulty is ever encountered in the Simon model in obtaining the needed social overhead capital – it drops like manna from heaven whenever the labor force grows. Policy makers who are interested in the process by which the social overhead capital comes into being may want to elaborate this portion of Simon's model. It is interesting to note before moving on that it is possible to interpret the relationship between the growth of the stock of social overhead capital and the growth of the labor force as a relationship between labor force growth and the pace of technological progress. If one believed that economies of scale due to the increasing specialization of the labor force occurred as the labor force increased in size, then the specification in equation (5.3) seems a bit more reasonable.

The capital stocks in the Simon model, as in the other models reviewed here, are determined by the cumulative addition of net investment to base year estimates of the values of the capital stocks. The determination of net investment by sector is discussed below. Given the indices of technology in the two sectors, the level of social overhead capital, and the capital stocks, the outputs of the sectors are determined once the labor inputs are known. In the Simon model, the labor inputs and sectoral outputs are determined simultaneously in a complex manner unique to this model. It is the explication of this mode of determining output that shall concern us for the next few pages.

5.2 SOCIAL INDIFFERENCE CURVES AND THE DETERMINATION OF AGGREGATE AND SECTORAL OUTPUT LEVELS

The Simon procedure for computing sectoral and aggregate output levels has three steps. First, the relative quantities of physical output of the two sectors in period t are postulated to depend upon income per consumer equivalent in period $t-1$. In symbols

$$\alpha(t) \equiv \frac{Q_I(t)}{Q_I(t)+Q_F(t)} = 0.35 + \left[\frac{\tilde{Y}(t-1)-75}{925}\right] \cdot 0.65 \tag{5.4}$$

where $\tilde{Y}(t-1)$ is income per consumer equivalent in period $t-1$ and $\alpha(t)$ is the proportion of total output in period t contributed by the industrial sector. Since Simon assumes that total output in period t, $Q(t)$, can be obtained by summing the physical quantities of outputs in the two sectors[27] [i.e., $Q(t) = Q_I(t) + Q_F(t)$], equation (5.4) may be rewritten using equations (5.1) and (5.2) as follows:

$$M_I(t) = \left[\frac{\alpha(t)\cdot Q(t)}{A_I(t)\cdot K_I^{0.4}(t)\cdot J(t)}\right]^{1/0.6} \tag{5.5}$$

and as

$$M_F(t) = \left[\frac{[1-\alpha(t)]\cdot Q(t)}{A_F(t)\cdot K_F^{0.5}(t)\cdot J(t)}\right]^{1/0.5} \tag{5.6}$$

Hence

$$M(t) \equiv M_I(t) + M_F(t) = \left[\frac{\alpha(t)\cdot Q(t)}{A_I(t)\cdot K_I^{0.4}J(t)}\right]^{1.67} + \left[\frac{[1-\alpha(t)]\cdot Q(t)}{A_F(t)\cdot K_F^{0.5}J(t)}\right]^{2.0} \tag{5.7}$$

Equation (5.7) provides Simon with a relationship between "aggregate output" $Q(t)$ and aggregate labor input $M(t)$.

One point on this output-labor frontier is chosen by society according to a social welfare mapping, which shifts around over time according to economic conditions. At any time t, Simon posits that we can write the jth member of the family of social indifference curves as follows:

$$\log\left[\frac{Q(t)}{L(t)}\right] = \alpha^*(t) + \beta_j^*\cdot\left[\frac{M(t)}{L(t)}\right] \tag{5.8}$$

where $L(t)$ is the total labor force in period t and β_j^* is a constant related to the index j,

$$\alpha^*(t) = \exp\{[0.4 - 0.2\cdot(\tilde{Y}(t-1)-75)/925]\cdot\tilde{Y}(t-1)\cdot C(t)/L(t)\} \tag{5.9}$$

and where $C(t)$ is the number of consumer equivalents in year t. The expression in equation 5.8 is supposed to capture the effects of relative aspirations, current standard of living, and the dependency ratio on social tastes for goods and leisure. In practice it may simply be said that $\alpha^*(t)$ depends upon the last period's per capita income and the current period's dependency rate. Given equations (5.8) and (5.7), the nation chooses a level of labor and output that maximizes its utility.

The determination of output via the process of maximizing a social welfare function is unique to the Simon model for good reason. Other model builders had in mind the ultimate objective of specific national applications of their models. This immediately rules out the Simon approach because of the impossibility of estimating the parameters of families of shifting social welfare functions. Simon, however, has built his model for the purpose of analysis, not ready applicability. But even for Simon's purposes, it is debatable whether the maximization of a social welfare function is the best framework to use. There can be no question on general grounds that one element of an economic-demographic simulation model should be the determination of the number of hours of work per labor force member per year. The conventional way of incorporating this into such a model would be to specify for each sector of the economy a supply of hours of work function that would relate hours of work supplied in the sector to the size of the sector's labor force, the dependency rate in the sector, the wage rate in the sector, and the nonlabor income (if any) accruing to workers in the sector. There is a substantial literature both theoretical and empirical to guide such a specification. There is, on the other hand, no literature that even suggests the existence, let alone the stability, of social welfare functions of the sort posited by Simon. Given the evidence at hand, prudence requires that the Simon social welfare function formulation be considered with an open, but a skeptical, mind.

One serious problem in the Simon model relates to the specification of net industrial investment. According to Simon, net industrial investment in period t may be written

$$NI_I(t) = 0.0275 \left[\log_{10} \left(\frac{Q_I(t) - Q_I(t-1)}{Q_I(t)} \right) \right] [1 - 0.5 \cdot YOU(t)] \cdot K_I(t) \tag{5.10}$$

where $NI_I(t)$ is net investment in the industrial sector in period t, $Q_I(t)$ is industrial output in period t, $YOU(t)$ is an index of the youth dependency burden in the entire country in period t, and $K_I(t)$ is the capital stock in the industrial sector in period t. The youth dependency burden is defined so as to be positive if the burden in year t is greater than in the base year and negative if the dependency burden is less than in the base year.

Clearly, this is a very odd specification for a number of reasons. First, net investment must always be *negative* except for extremely high values of the youth dependency rates. This occurs because $\log_{10}([Q_I(t) - Q_I(t-1)]/[Q_I(t)])$ is always negative when industrial output is growing. Further, the greater the youth dependency burden, other things being equal, the greater (less negative) is the quantity of net investment. This is exactly the reverse of the usual assumption that a greater dependency burden reduces capital formation. Is the specification in equation (5.10) an outright error that arose because Simon did not realize that the logarithm of a positive number less than unity is always negative? Perhaps. Possibly some other equation was used in the simulation program and the text is in error. Either alternative, however, suggests that extreme caution be exercised in interpreting any results from the Simon model.

5.3 TECHNOLOGICAL CHANGE

The same problem concerning the logarithm of a positive number less than one occurs in the specification of the rates of technological progress in the industrial and the agricultural sectors. In the base run Simon specified the rate of technological progress in the agricultural sector at one-half of one percent per annum. In symbols,

$$A_F(t+1) = 1.005 \cdot A_F(t) \tag{5.11}$$

In the industrial sector, the rate of technological progress was assumed to be lower than in the agricultural sector. The specification is

$$A_I(t+1) = Q_I(t)\left[1.005 + 0.002 \log_{10}\left(\frac{Q_I(t) - Q_I(t-1)}{Q_I(t)}\right)\right] \tag{5.12}$$

Since $\log_{10}([Q_I(t) - Q_I(t-1)]/Q_I)$ is a negative number, the rate of technological progress in the industrial sector in the base run is less than one-half of one percent per annum. Judicious modification of the parameters in equations (5.11) and (5.12) can easily allow technological progress to be more rapid in the industrial sector than in the agricultural sector, but no such results are reported in Simon's article.

5.4 DEMOGRAPHICS

There are no demographic specifications in the Simon model of any interest. Education is assumed to play no role in economic development. Labor force participation rates and fertility are assumed to be exogenous. Mortality rates are assumed to be a function of per capita income only – there are no public health expenditures in the model. Finally, migration does not depend on rural–urban income differences – such differences do not appear explicitly in the model – but rather adjust to whatever they need to be to make equation (5.4) true.

5.5 CONCLUSIONS

The Simon model, then, is not in its present form of much use to policy makers. Unusual formulations such as the assumption that net investment in the industrial sector is generally negative make the model grossly inapplicable to contemporary developing countries. Further, the specification that output and labor in any one period are determined so as to maximize a social welfare function is also problematical. The Simon framework, then, does not appear to be a useful one for further development. Policy makers interested in a more meaningful framework should begin with the Kelley–Williamson Representative Developing Country model described in Chapter 9.

6 THE FAO MODEL

The Food and Agriculture Organization's application of its systems simulation model to Pakistan is the simplest of the models reviewed here. Its simplicity is both its chief virtue and its chief defect, for while it is the easiest of all the models to implement, the FAO model is in many respects overly simplifed. This is unfortunate particularly because the FAO model is the only one of the group that purports to give serious guidance to agricultural policy makers.

6.1 AGRICULTURE

Eight productive sectors are incorporated into the FAO model: agriculture, small-scale industry, large-scale industry, capital goods industry, construction industry, traditional services, modern services, and government services. The agricultural sector itself is broken down into four subsectors: small-scale farming in rainfed regions, large-scale farming in rainfed regions, small-scale farming in irrigated regions, and large-scale farming in irrigated regions. Output growth in all sectors of the economy, including each of the agricultural subsectors, is assumed to be controlled by the government through its role in the allocation of investment funds.[28]

The government has a number of avenues for affecting agricultural production. It can consolidate small rainfed farms into large rainfed farms, consolidate small irrigated farms into large irrigated farms, decompose large irrigated farms into small irrigated farms, reclaim unused land for use in irrigated farming, invest in any of the four distinguished types of agriculture, and spend money on intermediate inputs. While this variety of agricultural policy instruments is certainly useful to agricultural planners, there are instruments omitted whose importance for agricultural planning are at least of equal consequence. In particular, the omission of all price variables from the FAO model means that no agricultural policy that affects agricultural output by affecting the relative price of farm produce can be considered.

The lack of any relative prices in a model of economic development poses serious problems, and these difficulties are magnified in a model that is to be useful for agricultural policy making. First of all, no change in the relative price of agricultural and industrial goods with economic development is allowed to occur in the model. To the extent that such a change does occur, the model is in error. Second, the model cannot be used to consider any agricultural pricing policies. For example, one might expect that a government subsidy to agriculture, say through the setting of a minimum sale price for important agricultural products, would, within a few years, cause the quantities of the subsidized commodities produced to increase. Further, resources might well be diverted from the production of the nonsubsidized products to the production of the subsidized ones. Yet no such effects of output pricing policies can be considered in the FAO model. Similarly, agricultural input pricing policies cannot be considered in the model. For example, there is no way of asking about the effects on agricultural output of a subsidy on fertilizer.

The FAO model is not unique in its assumption that all relative prices remain fixed forever. This assumption is made in three of the five second-generation models reviewed here. It is a poor assumption – one that is highly unlikely to approximate reality – and one potential problem area with all the models that incorporate it.

The agricultural policies that are allowed in the FAO model, unfortunately, are placed in such a simplified context that their operation does not appear to be closely linked with reality. In the FAO model, agricultural output is not related to agricultural inputs by a production function. Instead there is a set of land accounting equations and a set of equations determining yields per acre. The land accounting equations are straightforward. In the Pakistani simulations it is assumed that there is a fixed amount of land used in production in the rainfed regions. Small rainfed farms may be converted into large rainfed farms but not the reverse. Land in irrigated farming, on the other hand, is not assumed to be constant. Each year a certain amount of irrigated land is assumed to be withdrawn from cultivation, and a certain amount of irrigated land is, at a cost, reclaimed by the government. The net effect of these two forces may be either positive or negative. Both land consolidation and land distribution may occur in areas of irrigated farming.

Agricultural policies also are allowed to affect yields per acre. The expressions used to determine current yields have the form

$$Y(j,t) = Y(j,t-1) + \alpha_j \cdot \left[\frac{IN(j,t-1) - DIN(j,t-1)}{LA(j,t)}\right] + \beta_j \cdot IT(j,t) \tag{6.1}$$

where $Y(j, t)$ is the yield per acre on farms of type j in period t; $IN(j, t-1)$ is gross investment on farms of type j in period $t-1$; $DIN(j, t-1)$ is the cost of land consolidation, distribution, and reclamation on farms of type j in period $t-1$; $LA(j, t)$ is the amount of land used in the jth type of agriculture in year t; $IT(j, t)$ is the annual increment in the quantity per acre of intermediate inputs; and α_j and β_j are constants. The yield per acre on farms of the jth type in period t, then, depends upon the yield per acre of that type of agriculture in the previous period, net investment in that type of farming in the previous period, the amount of land used in the jth type of farming, and the quantity of intermediate inputs used in period t.

This specification of the determinants of agricultural productivity has a number of drawbacks. First, agricultural labor plays no role in producing output in the FAO model. It may be argued that agricultural labor is a redundant factor of production in many less developed countries today. But the assumption that labor will never attain a positive marginal product any time in the next thirty or so years regardless of the development strategy followed seems dubious at best. A second problem concerns the lack of capital depreciation in the FAO model. Investments in agriculture are unrealistically assumed to yield nondiminishing returns over the entire simulation period. Third, the specification assumes that lands whose status have altered immediately have the yields associated with the current agricultural type. In other words, if it is government policy to invest only in large consolidated farms in rainfed farming areas, such investment would raise the yield per acre on large consolidated farms. Further, if the government consolidated small holdings that had received no government investment, the yield per acre on the new consolidated farms still would equal the yield per acre on the consolidated farms on which investment took place. Since the cost of consolidating land (or distributing it) is fixed per acre regardless of yield differentials, the FAO model makes it appear as if changing the size of holdings provides the fruits of investment where none occurred.

A fourth sort of problem with the specification of the agricultural production arises because of the linearity of equation (6.1). There are three aspects of this difficulty that need to be discussed here – an obvious point and two somewhat more subtle ones. It is clear from inspecting equation (6.1) that there are no diminishing returns in the short run either to investment in any form of agriculture or to the incremental use of intermediate inputs. Thus, for example, the marginal yield gain per additional unit of fertilizer is assumed to be the same regardless of the level of incremental fertilizer use. It may be argued that in traditional agriculture the point of long-run diminishing returns to capital and intermediate inputs is so far in the future that it can safely be ignored in the simulations, but it is not clear that this argument is compelling with regard to diminishing returns in the short-run.

One somewhat less immediate result of the linearity of equation (6.1) concerns the relationship between incremental intermediate input use and the level of net agricultural output. The equation used in computing the latter is

$$ON(j, t) = OG(j, t) \cdot \left[\frac{ON(j, t-1)}{OG(j, t-1)}\right] - IT(j, t) \cdot LA(j, t) \qquad (6.2)$$

where $ON(j, t)$ is the net output of the jth type of agriculture in year t and $OG(j, t)$ is the gross output of the jth type of agriculture in year t.

It is quite likely that agricultural planners would use the FAO model to determine that incremental quantity of intermediate inputs in any year that would maximize net output. To see what advice the model would give them, multiply equation (6.1) by $LA(j, t)$ and substitute the resulting expression in place of $OG(j, t)$ in equation (6.2). This procedure produces the equation

$$ON(j, t) = k^* + IT(j, t) \cdot LA(j, t) \cdot \left[\beta_j \frac{ON(j, t-1)}{OG(j, t-1)} - 1\right] \tag{6.3}$$

where

$$k^* = [Y(j, t-1) \cdot LA(j, t) + \alpha_j \cdot IN(j, t-1) - \alpha_j \cdot DIN(j, t-1)] \cdot \left[\frac{ON(j, t-1)}{OG(j, t-1)}\right] \tag{6.4}$$

Clearly, if $\beta_j[ON(j, t-1)/OG(j-1)] - 1 < 0$, the net output of agriculture of the jth type is maximized in year t when $IT(j, t)$ is zero. If that expression is positive, net output is maximized when the incremental quantity of intermediate inputs is *infinite*! It should perhaps be noted in passing that unwary policy makers can be led significantly astray by this formulation. It certainly should be modified before serious analysis with the model is undertaken. One approach to mitigating this difficulty would be to assume that the costs of and returns from the use of intermediate inputs were not constant but rather varied with the quantities of those inputs consumed.

The third problem related to the linearity of equation (6.1) is closely akin to the one just analyzed. Suppose policy makers were to utilize the FAO model to determine the strategy that would maximize agricultural output[29] in a particular future year, given an exogenous annual series of total net agricultural investments. What advice would the model provide in such a situation? The answer is that, in general, to attain its goal the government should *at most* invest in *only one* of the four types of agriculture and *at most* in only one type of land conversion.[30] It is even possible that the government should spend its entire agricultural investment on a single activity. Thus, the linearity of equation (6.1) has a tendency to produce the implication that specialization is preferable to diversification.

Fortunately, the agricultural sector is embodied in a model that may help alleviate some of the specification's shortcomings. In the FAO model it is possible that the amount of investment in agricultural investment in a given year would depend in part on the level of agricultural output in previous years. In this case, the assumption made in the discussion above that the quantities of agricultural investment are exogenous does not hold, and the implications cited do not necessarily follow. Even though on purely technical grounds it is not possible to guarantee that the optimum agricultural policy involves specialization in investment, such a result is not an unlikely one. What, then, can an official ascertain about agricultural policy from experimenting with the FAO model? If his simulations suggest that the government strongly support only one or two types of agriculture, can he trust them? The answer, unfortunately, is that he should not. Such results are likely to arise because of the overly simplistic specification of the agricultural production process. If the simulations suggest that more should be spent on intermediate inputs like fertilizer, should he follow that suggestion? The answer, unfortunately, is uncertain. Net output is maximized by using either no additional amount of intermediate inputs or an infinite amount of them. In brief, the agricultural portion of the FAO model is too restrictive to be of much use in dealing with those questions it is designed to answer.

6.2 INDUSTRY

The nonagricultural portion of the FAO model is also quite simple. Seven nonagricultural sectors are distinguished in the model: small-scale industry, large-scale industry, capital goods industry, construction industry, small-scale services, large-scale services, and government services. Net output in each sector in year t depends upon net output in that sector in period $t-1$ plus the product of an exogenous amount of net investment and a fixed incremental output–capital ratio. The outputs in the six nongovernmental sectors are aggregated together by means of a set of invariant prices. Embodied technological progress may be introduced in the nonagricultural sector by systematically altering the incremental output–capital ratios, but no technological change is assumed to occur in the Pakistani simulations.

This specification of the determinants of nonagricultural production does have the advantage of being very easy to operationalize. It also shares the disadvantages discussed above in terms of the agricultural production relations. Further, omitting skilled and unskilled labor entirely from the nonagricultural production process involves implicit assumptions that hardly seem warranted, especially in a model that has a time horizon of several decades. It should also be noted here that demand conditions play no role whatsoever in the determination of output levels.

The implicit assumptions concerning constant returns to scale have been discussed above. In the nonagricultural portion of the economy, as opposed to the agricultural one, it is possible to prove that the government policy should direct investment toward only one nonagricultural sector, the one with the highest incremental income–capital ratio. To see this, it is necessary to note that net output in nonagricultural sector j at time t years after the beginning the simulation is simply

$$ON(j, t) = ON(j, 0) + \sum_{\tau=1}^{t} IN(j, \tau) \cdot \kappa(j), \tag{6.5}$$

where $ON(j, t)$ is the net output of the jth sector t years after the beginning of the simulation, $IN(j, \tau)$ is investment in the jth sector in year τ, and $\kappa(j)$ is the incremental output–capital ratio in sector j. Further, since all relative prices are fixed at unity, aggregate nonagricultural output in year t, $\overline{ON}(t)$, may be written

$$\overline{ON}(t) = \sum_{j=1}^{6} ON(j, 0) + \sum_{j=1}^{6} \sum_{\tau=1}^{t} IN(j, \tau) \cdot \kappa(j) \tag{6.6}$$

Given any amount of investment in the nonagricultural sector, it is clear that aggregate nonagricultural output in *every* year of the simulation period is maximized by investing only in that sector with the highest marginal product of capital. Since nonagricultural output is maximized in every year by investing in only one sector, this strategy will be the one chosen to meet any policy goal. Policy makers experimenting with the FAO model as applied to Pakistan will find that economic growth will proceed fastest when the government policy induces investment only in small-scale industry and, since it has an identical incremental capital-output ratio, large-scale modern services.

The linear output specifications in the FAO model build in an important conclusion about whether developing countries should concentrate their resources in encouraging agricultural or industrial growth. To answer this question in the context of the FAO model is reasonably straightforward. It translates into asking whether the rate of return on investment in its most productive agricultural use is greater or smaller than it is in its most productive nonagricultural use. Let us consider how this question is answered in the Pakistani case. In small-scale industry and the modern service sector the rate of return on investment is 33 percent. These are the highest rates of return available in the nonagricultural sector with the exception of the construction industry.[31] The rate of return on an investment in rainfed agriculture, on the other hand, holding the stock of land in rainfed agriculture constant, is 80 percent per annum.[32] Clearly, development should be based on rainfed agriculture and not on industry. Indeed, the optimum development strategy in the FAO model is to spend nothing on industrial growth.

It should be noted in passing that a country that can invest substantial amounts of money at rates of return in the neighborhood of 80 percent per annum without the risk of diminishing returns should without much strain be able to enjoy stupendous rates of economic growth. Indeed, a policy maker experimenting with the FAO model will soon discover that the secret of achieving spectacularly high sustained rates of economic growth is simply to invest all the government's funds in rainfed agriculture or, if he prefers a more balanced development strategy, in large farms in rainfed regions, small-scale industry, and modern services.

6.3 FINAL DEMAND

The FAO model does not deal with factor payments of any kind. Therefore, policies that affect demographic or economic variables through changes in wage rates, profits, or rents cannot be analyzed in the context of the model.

This neglect of factor payments is related in a formal way to the FAO model's neglect of relative output prices. Since factor payments do not appear in the model, any effects arising from changes in the functional distribution of income cannot be studied.

In the FAO model, the entire income side of the national income accounts is ignored. Per capita private consumption in period t is assumed to be equal to the product of per capita private consumption in period $t - 1$ and a multiplier that depends upon the rate of growth of per capita income. It is stipulated in the FAO model that per capita private consumption can never decline. Government consumption grows each year by an amount determined by the product of the amount of money the government invested in itself in the previous year and a constant incremental consumption–investment coefficient.

Investment (net and gross because there is no depreciation) is defined to be equal to the value of gross domestic product minus private and governmental consumption plus net imports. The value of net imports in the FAO model is considered to be a policy variable set by the government, so net investment is known once aggregate output and total consumption are determined. All investment funds are assumed to be allocated according to exogenous policy rules. No mention is made of whether the fixed rates of return to capital are used in the allocation decisions.

A dollar invested in any of the sectors in year t is assumed to result in a fixed derived demand for the output of the construction industry.[33] Further, since a fixed proportion of the output of the construction industry is to be used for purposes other than net investment, it is clear that there will generally be either excess demand or excess supply in the construction industry.[34] To solve this problem, which typically arises in fixed-price models when elements of both the demand and supply side are considered, the FAO model introduces an *ad hoc* adjustment, which unfortunately does not always perform its intended function.

The adjustment works in the following manner. If in any year either (a) the derived demand for construction exceeds the supply of construction output available to meet that demand or (b) the supply exceeds the demand by some predetermined amount, then the investment allocation to construction in the *previous* year is altered. Further, the investment allocation to every other sector of the economy in the *previous* year must be modified in order to keep total investment constant. This process of reallocating investment allocations only refers to the year prior to the current one. A regression in this manner back to the first year of the simulation is explicitly forbidden. The object of this *ad hoc* procedure is, it appears, to ensure that the difference between the derived demand for construction and the supply available to meet that demand is small and nonpositive. Generally, this *ad hoc* adjustment will not yield the desired result except in the last year of the simulation. Worse still, there is a set of conditions that a policy maker may encounter while experimenting with the FAO model under which the adjustment procedure completely breaks down.

Let me support these two assertions with some simple analysis. First, let us assume that, by adjusting the investment allocations in year $t-2$, the construction industry is in equilibrium in year $t-1$. Now, let there initially be excess demand for the output of the construction industry in period t. To eliminate the excess demand in period t, investment allocations in period $t-1$ must be altered in favor of the construction industry. But before this alteration of investment flows the construction industry was in equilibrium! Generally, these changes in investment patterns will cause the construction industry, which in period $t-1$ had neither significant excess demand or supply, to develop one or the other. Thus, the construction industry adjustment for period t causes the construction industry in period $t-1$ to be out of equilibrium, the construction industry adjustment for period $t+1$ causes the construction industry in period t to be in disequilibrium, and so on until finally the only year in which the construction industry is in equilibrium is the last one in the simulation period.

As strange as this adjustment process now must appear, it has an even worse feature – it can break down entirely. Let us begin again in the situation in which the construction industry is in equilibrium in period $t-1$ but initially in a state of excess demand in period t. Clearly, we must return to period $t-1$ and allocate more money to investment in the construction industry, and this money must be taken away from investments in other sectors. It is possible, however, that further investment in construction in period $t-1$ will result in an *increase* in the derived demand for construction in period $t-1$.[35] But this increase in demand cannot be met with the capacity on hand in period $t-1$! Thus, it may be impossible to reallocate funds in period $t-1$ to meet an incipient situation of excess demand in period t. What happens to the FAO model when such a situation occurs is not discussed. Policy makers nonetheless should be aware of this problem.

The FAO model does contain a few equations on foreign trade. The major assumption there is that the balance-of-payments deficit, or, equivalently, the balance-of-trade deficit – there are no capital flows in the model – is exogenously determined by the government through its control over exports. The equations make no mention of the country's exchange rate or of a long-run balance-of-payments constraint.

6.4 EMPLOYMENT

Although employment has no effect on output in the FAO model, output growth does influence the growth of employment in large-scale industry, construction, capital goods production, and large-scale modern services (excluding the government). Increases in employment in any of those sectors is posited to be determined by the product of the increase in sectoral output and a sector-specific incremental employment–output ratio, defined as the change in employment divided by the change in output. These ratios are not held constant, but rather change according to a ratchet-type mechanism. In order to understand how the incremental employment-output coefficients vary, let us define $e(j, t)$ to be the incremental employment-output coefficient for industry j in period t. The equation determining $e(j, t)$ may be written

$$e(j, t) = e(j, t-1) \cdot \{1 + \min(0, \beta(j) \cdot [u(t-1) - u(t-2)])\} \quad (6.7)$$

where $\beta(j)$ is a positive constant specific to sector j and $u(t-1)$ is the unemployment rate in the large-scale modern sectors[36] in period $t-1$.

Equation (6.7) says that if the unemployment rate in the modern large-scale sectors drops by one percentage point, say from 10 to 9 percent from period $t-2$ to period $t-1$, then the incremental employment-output ratio in period t will be smaller than its value in period $t-1$ by $\beta(j)$ percent. If, alternatively, the unemployment rate in the modern large-scale sectors increases from period $t-2$ to $t-1$, then the incremental employment-output ratio in period t will be unchanged from its previous period's value. In brief, increases in the unemployment rate do not affect the incremental employment-output ratios, while decreases in the unemployment rate cause those ratios to decline. If the unemployment rate had a tendency to move cyclically around a constant trend, the $e(j, t)$ would have a tendency to continue declining until their low values caused the unemployment rate *in the model* to begin a secular increase. The high predicted unemployment rates in the Pakistani simulations, however, cannot be attributed to this mechanism, since in those simulations the $\beta(j)$ were all set equal to zero.

Regardless of whether the $\beta(j)$ are set equal to zero or not, the relationship between capital, labor, and output would be much more plausible if some production function were consistently used. In that framework it is much easier to formalize the concept of the proximate determinants of the quantity of labor demanded.

6.5 LABOR FORCE

The aggregate labor force in the FAO model is determined by weighting the entire population by a set of constant age- and sex-specific labor force participation rates. Neither the possibility that labor force participation rates could vary over time as economic development occurs nor the possibility that labor force participation rates can vary by rural or urban residence is discussed. The growth of the aggregate labor force, then, is determined by purely demographic factors. In order to define the unemployment rate in the modern large-scale sectors of the economy, the labor force in these sectors must be defined. Conceptually this is not a straightforward task because it is unclear whether the labor force in the modern large-scale portion of the economy should be considered to be the entire urban labor force or whether a more restricted definition should be used. In practice, however, this problem disappears. Labor force surveys yield data on employees in modern large-scale industries and on all people seeking jobs but not currently employed. This combination is taken to be the base-year observation on the size of the labor force associated with modern large-scale industries.

Subsequent to the base year, it is assumed that the labor force associated with modern large-scale industries has two sources of growth: natural increase and transfers from the remainder of the labor force. The natural increase of this modern labor force is assumed to be identical to the rate of increase of the aggregate labor force. It is possible to argue that the "natural" rate of growth of the modern labor force is likely to be lower than the rate of growth of the aggregate labor force, because the former is more urban and more educated than the latter. The magnitude of any error introduced by that assumption, however, will be trivial relative to the other problems in the model.

The specification of the number of people transferring to the modern labor force from the remainder of the labor force is given in equation (6.8)

$$TR(t) = TR(t-1)\left[\frac{LFR(t)}{LFR(t-1)}\right]\cdot\left[\frac{GR(t)}{GR(t-1)}\right]\cdot PD(t) \qquad (6.8)$$

where $TR(t)$ is the number of people transferring to the modern labor force in period t, $LFR(t)$ is the number of people in the residual labor force in period t, $GR(t)$ is a gravity constant for period t whose role in this equation is discussed below, and $PD(t)$ is a constant that depends upon the relative growth rates of the output per labor force member in the modern large-scale sectors compared with that in the remainder of the economy.

Another way of viewing this is to rewrite equation (6.8) as

$$\frac{TR(t)}{LFR(t)} = \left[\frac{TR(t-1)}{LFR(t-1)}\right]\cdot\left[\frac{GR(t)}{GR(t-1)}\right]\cdot PD(t) \qquad (6.9)$$

Recursively substituting the expression for the transfer rate in equation (6.9) into the right-hand side of that expression yields

$$\frac{TR(t)}{LFR(t)} = \left[\frac{TR(0)}{LFR(0)}\right]\cdot\left[\frac{GR(t)}{GR(0)}\right]\prod_{\tau=0}^{t} PD(\tau) \qquad (6.10)$$

Thus, the current transfer rate depends upon the transfer rate at the beginning of the simulation period, the gravity constant in period t relative to its value at the beginning of the simulation period, and the product of all the $PD(\tau)$ from the beginning of the simulation period up through year t.

In the FAO model, the gravity multiplier is defined by the following equation

$$GR(t) = \sigma_M(t)\cdot[1-\sigma_M(t)] \qquad (6.11)$$

where $\sigma_M(t)$ is the fraction of the total labor force in the modern sectors in year t. Clearly, $GR(t)$ is a symmetric function of $\sigma_M(t)$ over the interval [0, 1] that reaches a maximum at $\sigma_M(t) = 0.5$. The rate of transfer then increases, other things constant, as $\sigma_M(t)$ becomes closer to one-half, and decreases as it deviates more from that figure. Whether this assumption is generally accurate remains to be demonstrated. A policy maker using the FAO model should check the plausibility of the specification of the gravity multiplier for his own country.

The productivity differential term $PD(t)$ is computed using the following expression

$$PD(t) = 1 + \gamma \cdot \left[\frac{r_M(t)}{r_r(t)} - 1\right] \tag{6.12}$$

where $r_M(t)$ is the rate of growth over the previous period of output per labor force member in the modern sectors (excluding the government), $r_r(t)$ is the rate of growth over the previous period of output per labor force member in the remainder of the economy, and γ is a positive constant.

6.6 THE DEMOGRAPHICS

The demographic portion of the FAO model was not implemented in the Pakistani case because of lack of data. A family of population projections was used in its place. The following comments on the demographic specification are based on the prototype model (see pp. 100–104 of FAO 1976). The basic population accounting system can be improved. It does not maintain any information by single years of age and thus cannot age the population in a straightforward manner by applying single-year-of-age survival rates. The use of age-aggregated data makes the demographic accounting less precise than it would be if the simpler alternative of maintaining the age detail were followed. The impact of this imprecision, however, will be quite small in general.

The education accounting equations are similar to the demographic accounting equations. There is no behavioral content in either set. Educational policy can be seriously treated in the FAO model only after careful consideration is given to how education affects other variables in the model, for example, labor productivity and rural–urban migration.

The basic fertility variable in the prototype model is the general fertility rate.[37] The basic equation determining the general fertility rate is

$$\frac{GFR(t)}{GFR(t-1)} = 1 - a_1 \cdot ED(t-1) - a_2 \cdot JF(t-1) \tag{6.13}$$

where $GFR(t)$ is the general fertility rate in time t, $ED(t-1)$ is a term related to the average educational level of adults in period $t-1$, $JF(t-1)$ is a rough proxy for the rate of change of job opportunities for women in the modern sectors between period $t-2$ and period $t-1$, and a_1 and a_2 are positive constants. The precise definitions of $ED(t-1)$ and $JF(t-1)$ are given below. Before they are discussed, however, two aspects of equation (6.13) deserve attention. First, it should be noted that the process of urbanization is assumed to have no impact on fertility levels. Any policy maker who uses this equation should check to see if this is an appropriate assumption for his country. Second, holding $ED(t-1)$ and $JF(t-1)$ constant, the rate of change in the general fertility rate is constant. If that rate of change is positive, the general fertility rate continues to increase indefinitely and in the limit approaches positive infinity. If that rate of change is negative, the general fertility rate continues to decrease indefinitely and in the limit approaches zero. The implausibility of these inferences suggests that the relationship between the general fertility rate and its determinants ought in future work to be made more realistic.

The variable $ED(t)$ is defined by the following equation:

$$ED(t) = \max\,[EA(t), EA(t)\cdot\phi(t)] \tag{6.14}$$

where $EA(t)$ is the average adult level of education in period t and $\phi(t)$ is a population policy multiplier. In the FAO model there is no cost associated with changing $\phi(t)$, and thus the government can always obtain any general fertility rate it wishes by choosing an appropriate level of $\phi(t)$. Population policy is vastly more complex than this. It is clear, on this account alone, that serious work concerning population policy cannot be done in the context of the FAO prototype model.

The variable $JF(t)$ in equation (6.13) is supposed to be closely related to the rate of change of job opportunities for women in the modern sectors. The equation defining this variable is

$$JF(t) = \max\,[0, \rho_M(t) - \rho_F(t)] \tag{6.15}$$

where $\rho_M(t)$ is the rate of growth between period $t-1$ and t of employment in the modern sectors and $\rho_F(t)$ is the rate of growth of the number of females in the reproductive ages in the population as a whole between period $t-1$ and period t. The difference between the two growth rates is not unambiguously a measure of the job opportunities for women, since the proportion of women in the reproductive ages who can take advantage of job openings in the modern sectors is likely to change over time. Further, $\rho_M(t)$ can rise, but if the number of males seeking the new jobs rises even faster, opportunities for women may even decline. In addition, it is not clear why, if $\rho_M(t) - \rho_F(t) < 0$ implies a decline in fertility (relative to the situation where $\rho_M(t) - \rho_F(t) = 0$), then $\rho_M(t) - \rho_F(t) > 0$ does not imply a relative increase in fertility.

Mortality rates in the prototype model are to be generated from a model life-table system, given a value of the life expectancy at birth. The trend in this life expectancy may be determined either exogenously or endogenously given per capita consumption and government service investment. Policy makers should be warned that the endogenous determination of life expectancy in the FAO model may be inappropriate for their countries.

The rural–urban migration process is identical with the sectoral switching process discussed above except that the residual labor force is replaced by the rural population and the modern labor force is replaced by the urban population. With the appropriate modifications, the comments made above about the switching process apply as well to the specification of urban–rural migration.

6.7 CONCLUSION

In summary, the FAO model, although it is simple to implement, suffers from the disadvantages of that virtue. The linearity of the production relationships, the elimination of labor's role as a determinant of output levels, the lack of any capital depreciation, the absence of any demand structure, the lack of attention to the distribution of income (among other things), all strongly suggest that the policy prescriptions of the FAO model be treated very cautiously.

7 THE KELLEY, WILLIAMSON, AND CHEETHAM MODEL

Of the five second-generation models, the earliest one is the Kelley, Williamson, and Cheetham (KWC) model of dualistic economic development in Japan from the mid-1880s to the First World War. In addition to being the earliest of the second-generation economic–demographic simulation models, the KWC model provides the best framework for policy analysis among all of them.

7.1 THE RELATIONSHIPS BETWEEN INPUTS AND OUTPUTS

The KWC model recognizes two sectors of the economy: an agricultural sector and an industrial sector. The former is considered to be entirely rural, while the latter is assumed to be entirely urban. The functions relating inputs to outputs in the two sectors are restricted constant elasticity of substitution production functions. In the industrial sector, the production function may be written

$$y_I(t) = A_I\{[e^{\lambda_K t} K_I(t)]^{\rho_I} + [e^{\lambda_L t} L_I(t)]^{\rho_I}\}^{1/\rho_I} \qquad (7.1)$$

where $y_I(t)$ is the number of physical units of industrial output in period t, $K_I(t)$ is the capital stock in the industrial sector in period t, $L_I(t)$ is employment in the industrial sector in period t, λ_K is the rate of capital-augmenting technological progress in the industrial sector, λ_L is the rate of labor-augmenting technological progress in the industrial sector, ρ_I is a constant related to the elasticity of substitution between labor and capital in the industrial sector,[38] and A_I is a constant. The production function for agricultural output is analogous to the industrial production function and may be written

$$y_A(t) = A_A\{[e^{\mu_K t} K_A(t)]^{\rho_A} + [e^{\mu_L t} L_A(t)]^{\rho_A}\}^{1/\rho_A} \qquad (7.2)$$

where all the variables and parameters are defined like those in the industrial production function except that they all refer to agriculture.

Since these production functions are among the key elements of the KWC model, it is useful to discuss them in some detail. These constant elasticity of substitution production functions are the most sophisticated production functions used in any of the models reviewed here. This production structure has the advantage that it allows the elasticity of substitution between capital and labor to be different in the two sectors of the economy. It has a further advantage that differential rates of factor-augmenting technological progress may occur for a given factor across sectors and for the factors in a given sector. Indeed, an important element in the analysis of Japanese economic development in the KWC model is the sectoral difference in the bias of technological change. Such a phenomenon cannot be captured in any of the other production structures.

Both production functions assume constant returns to scale in any period. Further, it is assumed implicitly that agricultural production requires no inputs from the industrial sector (except agricultural capital) and that industrial production requires no raw materials from the agricultural sector. Someone interested in agricultural policy questions may want to modify these two assumptions. In particular, inputs from the modern sector such as fertilizer and electricity should be allowed to play a role in agricultural production. Similarly, agricultural inputs into industrial production should be allowed, if only to represent food processing. It should be noted that land does not explicitly appear in the agricultural production process. To the extent that land policy is important in a particular case the KWC model would have to be modified to reflect that.

As general as the KWC production structure appears, it does have one relatively subtle difficulty of which policy makers should be aware. The CES production functions in the KWC model are restricted in a special way – and this restriction has important implications for the interpretations of the CES parameters. A general two-input CES production function can be written:

$$y = A\{\delta K^{\rho} + (1-\delta)L^{\rho}\}^{1/\rho} \tag{7.3}$$

where y is output, K is capital, L is employment, and A, δ, and ρ are constants. In the KWC production functions, the constant δ does not appear. The disappearance of that parameter implies that $\delta = 0.5$ and that its effect is captured in the constant term A. This is an extremely rigid restriction to put on a CES production function. Among other things, it implies that if the elasticity of substitution is close to unity, then the factor shares must be close to 50 percent and, conversely, if the factor shares do not approximate one-half, the elasticity of substitution cannot approximate unity.

Real-world data, however, may well be generated by a production process that has factor shares nowhere near one-half, but that still has an elasticity of substitution approximating unity. To see what effect such a situation would have, we performed the following conceptual experiment. Hypothetical data were generated by a Cobb–Douglas production function where labor's share was 75 percent, capital's share was 25 percent, and there was no technological progress. A CES production function of the type used in the KWC model was then fitted to these data. The result was that it was possible to produce with such data CES parameter estimates which indicated (a) an elasticity of substitution considerably below unity and a labor-saving bias in technological change and (b) an elasticity of substitution considerably above unity and a labor-using bias in technological change. These configurations are the assumptions made for industry and agriculture respectively in the KWC model. Thus, policy makers should be cautious about statements made concerning elasticities of substitution and biases in the rates of factor-augmenting technological progress on the basis of CES production functions from which the distribution parameter δ is absent.

7.2 THE DISTRIBUTION OF INCOME, SAVINGS, AND CONSUMPTION

Payments to the four factors of production in the KWC model are made according to the values of their marginal products. The functional distribution of income, as we shall see below, plays an important role in determining the aggregate saving rate in the economy. It may be argued by some that the neoclassical assumption that factors of production are paid the values of their marginal products does not hold in contemporary less developed countries and that therefore the KWC approach ought to be abandoned. Although the premise of this argument may certainly be true, the conclusion hardly follows from it. Distortions in factor markets can easily be introduced into the KWC framework. Indeed, one addition to the KWC model that policy makers may wish to make is to formalize the factor market distortions that they believe to be most important in their own countries.

Given a sensible functional distribution of income, it is relatively easy to progress to a plausible specification of savings behavior. In the KWC model, the simplest possible saving equations are introduced. It is assumed that there is no saving out of labor income and that a fixed proportion of income from capital is saved. It is possible, of course, to envision a more complex specification of the determinants of savings, and, indeed, such an addition may be useful in the context of policy analysis for certain countries.

Given the functional distribution of income in the economy and the relative prices of industrial and agricultural goods, the KWC model determines the demands for those goods using a modified Stone–Geary system of demand equations. In this aspect of model building the KWC model towers above the others discussed here. The KWC model is the only one in which the prices of goods play a plausible role in influencing the quantities of goods demanded. There are six basic consumption demand equations in the KWC model:

$$D_{Ij}^{L}(t) = \frac{\beta_{Ij}[W_j(t) - \delta] \cdot L_j(t)}{P(t)} \qquad (j = I, A) \tag{7.4}$$

$$D_{Aj}^{L}(t) = [\beta_{Aj}W_j(t) + (1 - \beta_{Aj})\delta] \cdot L_j(t) \qquad (j = I, A) \tag{7.5}$$

$$D_I^K = \frac{\Pi_I[(1-S)k(t)-\delta]}{P(t)} \cdot \hat{K}(t) \tag{7.6}$$

$$D_A^K = [\Pi_A(1-S)k(t) + (1-\Pi_A)\delta] \cdot \hat{K}(t) \tag{7.7}$$

where $D_{ij}^L(t)$ is the demand for the goods of sector i by employed workers in sector j in year t, D_i^K is the demand for the goods of sector i out of capital income received in period t, $W_j(t)$ is the per worker labor income of people employed in sector j in period t, $L_j(t)$ is the number of people employed in sector j in year t, $P(t)$ is the ratio of the price of industrial goods to the price of agricultural products, S is the savings rate out of income from capital, $k(t)$ is the average amount of capital income per recipient of capital income in year t, $\hat{K}(t)$ is the number of recipients of capital income in period t, and $\beta_{II}, \beta_{IA}, \beta_{AI}, \beta_{AA}, \Pi_I, \Pi_A$, and δ are constants.

It is not necessary to discuss the properties of the Stone–Geary system of demand equations here. There are, however, two points worth mentioning briefly. First, the Stone–Geary system is quite flexible. With only minor modifications in the equations it is likely that a policy maker can specify a system of demand relations that is appropriate for his country. Second, since the constants in the demand functions differ by income type, changes in the functional distribution of income alter both the aggregate savings rate and the pattern of demand. The impact of these differential consumption patterns on the pace and character of the development process may be quite important, and they should not be overlooked by policy makers or model builders.

7.3 GENERAL EQUILIBRIUM CONSIDERATIONS

Given the functional distribution of income and the relative price of industrial goods, savings and the consumption demands for the economy's two products are determined. Since it is postulated that all savings are invested and that all investment is manifested by a demand for the industrial good, these conditions determine the vector of final demand.[39] The relative price ratio $P(t)$ is computed so that the output of each of the two sectors exactly equals the quantities of those products demanded. The KWC model, then, is, technically speaking, a general equilibrium model in which the relative price ratio, output levels, consumption, investment, and functional distribution of income are all determined simultaneously.

The advantages for policy analysis of having a general equilibrium framework, even if there are distortions, are numerous. A model in which the terms of trade between industry and agriculture are endogenous allows a policy maker to analyze decisions whose primary impact is on those terms of trade. Endogenous factor incomes allow policy makers to consider the effects of policies that primarily affect various income flows. Indeed, in the framework of the KWC models one can determine what the effect will be on relative output prices of a government's attempt at changing consumers' purchasing patterns. When both the supply and demand sides of the economy are allowed to interact properly in a model, it is much easier to use that framework to pose and answer policy questions than if only the demand or only the supply side of the economy is present in the model. It is the successful integration of the supply and demand sides of the economy that sets the KWC model apart from the other second-generation models studied here and that makes it a good foundation on which to add further developments.

7.4 DYNAMIC ASPECTS

Several dynamic aspects of the KWC model remain to be discussed. Of particular importance are the problems of allocating investment expenditures across sectors and determining the volume of rural–urban migration. The formal specifications of both these processes are identical in KWC models, so for convenience they will be discussed together. For each of those two facets of the model, an equilibrium and a disequilibrium formulation are given. The equilibrium specification of the investment allocation problem begins with the assumption of costless capital mobility. That assumption implies that the value of the aggregate capital stock in the country plus the amount of investment in the current year is treated as an annual flow variable that is allocated to the two sectors so as to equalize the rate of return on capital across the sectors in each year. The equilibrium formulation of the migration problem starts with the assumption of costless migration. In this case, the inference is that the labor force divides itself across sectors so as to equalize wage rates in the two sectors.[40] Neither of these equilibrium formulations, however, is very persuasive.

In reality, neither capital mobility nor labor mobility is perfectly costless. In order to represent formally the kind of imperfect capital and labor mobility that occurs in reality, the KWC model provides two disequilibrium formulations. Capital mobility in this latter view is allowed only in the allocation of current investment funds. Capital, once put in place, is considered forever immobile. The total amount of money invested in each sector depends upon the distribution of savings by sector of origin and upon the relative rates of return in the two sectors. The basic equations of the disequilibrium framework are

$$S_{II}(t) = \$_I(t) \quad \text{if} \quad r_A(t) - r_I(t) < \tau \qquad (7.8)$$

$$S_{II}(t) = \$_I(t)\, e^{\mu[r_I(t) - r_A(t) + \tau]} \quad \text{if} \quad r_A(t) - r_I(t) \geqslant \tau \qquad (7.9)$$

$$S_{IA}(t) = 0 \quad \text{if} \quad r_A(t) - r_I(t) < \tau \qquad (7.10)$$

$$S_{IA}(t) = \$_I[1 - e^{\mu[r_I(t) - r_A(t) + \tau]}] \quad \text{if} \quad r_A(t) - r_I(t) \geqslant \tau \qquad (7.11)$$

$$S_{AA}(t) = \$_A(t) \quad \text{if} \quad r_I - r_A(t) < \tau \qquad (7.12)$$

$$S_{AA}(t) = \$_A(t)\, e^{\mu[r_A(t) - r_I(t) + \tau]} \quad \text{if} \quad r_I(t) - r_A(t) \geqslant \tau \qquad (7.13)$$

$$S_{AI}(t) = 0 \quad \text{if} \quad r_I(t) - r_A(t) < \tau \qquad (7.14)$$

$$S_{AI}(t) = \$_A(t) \cdot [1 - e^{\mu[r_A(t) - r_I(t) + \tau]}] \quad \text{if} \quad r_I(t) - r_A(t) \geqslant \tau \qquad (7.15)$$

where $S_{ij}(t)$ is the savings generated in sector i invested in sector j in time period t, $\$_i(t)$ is the total savings generated in sector i in period t, $r_i(t)$ is the rate of return on capital in sector i earned in period t, and μ and τ are constants that can be affected by governmental policies.

Although this specification appears rather cumbersome, it is truly quite simple. Since the explication is identical for investment generated in each sector, it will be sufficient to discuss only investment in the industrial sector. All investment generated in the industrial sector is assumed to be invested in the industrial sector unless there is a rate-of-return differential favoring agriculture of at least τ percentage points. As the rate-of-return differential favoring agriculture grows larger, the fraction of urban savings invested in the rural area grows larger and asymptotically approaches unity as the differential approaches infinity. This is a plausible representation of the allocation of investment funds even where capital markets are poorly developed.

The disequilibrium formulation of the migration process works in much the same manner. The motivating force behind rural–urban migration is the expected income differential between urban and rural areas. The rate of rural-to-urban migration is assumed to be

$$m(t) = 1 - e^{\rho w^*(t)} \qquad (7.16)$$

where $m(t)$ is the rate of rural–urban migration in year t, $w^*(t)$ is rural–urban income differential adjusted for the costs of migration and ρ is a constant.[41]

Given the sectoral allocation of investment and the determination of rural–urban migration, there remains only one dynamic element of the model left to discuss – the rate of growth of the labor force. In the KWC model, the rates of growth of the industrial labor force and the agricultural labor force are exogenous parameters. Thus, except for migration, the KWC model does not allow for any influences running from the economy to the demography of the country. Policy makers interested in a full-scale demographic–economic simulation model will have to supplement the KWC model here with formulations that are relevant to their country.

7.5 CONCLUSION

The KWC model, in its present form, is strong economically but underdeveloped demographically. This is clearly appropriate for the purposes of the model builders, but it is inappropriate from the perspective of those interested in economic–demographic interrelationships. Agricultural policy makers in particular will find that there is much of interest that can and should be incorporated into the KWC framework in order to make it useful for them.

8 THE ADELMAN-ROBINSON MODEL OF KOREA

It is useful to consider here two third-generation models, the Adelman-Robinson model of Korea and the Kelley-Williamson model of a representative open-economy developing country. Neither of these models has a well-articulated demographic aspect, and therefore they do not technically belong in a review of economic-demographic simulation models. It is useful, however, to investigate their structures, because it will be on frameworks such as these that the third generation of economic-demographic simulation models will be constructed. Reviewing these two models, then, allows us a glance into the future.

The Adelman-Robinson simulation model of the Korean economy differs from the second-generation economic-demographic simulation models reviewed above in that it has a medium-term focus. The simulation period is never allowed to be longer than 9 years. As a consequence of this focus many of the economic-demographic linkages highlighted in the other models are omitted from this one. The Adelman-Robinson model also differs from the other models reviewed here in its detailed consideration of the country's financial and monetary structures. These differences are quite significant and make comparison of the Adelman-Robinson model with the others somewhat difficult. The central question addressed by the Adelman-Robinson model, however, is the same as that addressed by the Bachue model, the relationship between economic growth and the distribution of income. Therefore, it will be useful to ascertain how two quite different models approach the same problem.

The Adelman-Robinson specification is divided into three stages. The effects of the financial structure of the Korean economy on the allocation of nominal investment funds are determined in stage I. These allocations are allowed to depend on expectations of future sales and prices, which may or may not be subsequently realized. Stage II is a static general equilibrium model that takes the results of stage I as given. This portion of the model not only determines relative prices endogenously but also determines the rate of inflation. The third stage is composed of dynamic equations that take the results of the second stage and update endogenous variables so that the model can return to stage I. In the presentation of the model below, we shall discuss stage II first, and then stages III and I.

8.1 PRODUCTION RELATIONS

The Adelman-Robinson model differentiates between 29 sectors of the Korean economy: rice, barley, and wheat production; other agricultural output; fishing; processed foods; mining; textiles; finished textile products; lumber and plywood; wood products and furniture; basic chemical products; other chemical products; petroleum products; coal products; cement; nonmetallic and mineral products; metal products; nonelectrical machinery; electrical machinery; transport equipment; beverages and tobacco; other consumer products; construction; electricity and water; real estate; transportation and communications; trade and banking; education; medical services and other services; and personal services. In each of these 29 sectors, the model delineates four firm (farm) sizes; thus it requires 29 x 4 or 116 separate production formulations.

Two types of production functions are used in the model, Cobb-Douglas and two-level CES. The Cobb-Douglas specification, used in the 18 nonfarm, nonservice sectors, is

$$X_{is}(t) = A_{is}(t) K_{is}^{\alpha_{is}}(t) \prod_{\lambda=1}^{n_{is}} L_{is\lambda}^{\beta_{is\lambda}}(t) \tag{8.1}$$

where $X_{is}(t)$ is the physical output of firms of size s in sector i in period t, $A_{is}(t)$ is the productivity constant for firms of size s in sector i in period t, $K_{is}(t)$ is the relevant capital stock, and $L_{is\lambda}(t)$ is the amount of labor of skill type λ employed in firms of size s in sector i in period t, and the parameters $\alpha_{is}, \beta_{is1}, \beta_{is2}, \ldots$ sum to unity, and n_{is} is the number of labor skill types employed by firms of size s in sector i.

Output in the two-farm sectors is modeled by two-level CES production functions of the form

$$X_{is}(t) = A_{is}(t)[\alpha_{is} \cdot L_{is}^{-\rho_i}(t) + (1-\alpha_{is})K_{is}^{-\rho_i}(t)]^{-\gamma_{is}/\rho_{is}} \tag{8.2}$$

where

$$L_{is}(t) = k \prod_{\lambda=1}^{n_{is}} L_{is\lambda}^{\beta_{is\lambda}}(t) \tag{8.3}$$

and where $X_{is}(t)$ is the output of farms of size s in sector i in period t, $A_{is}(t)$ is the relevant productivity constant, α_{is} is the CES distribution parameter for farms of the (i, s) type, $L_{is}(t)$ is the aggregate labor input measure formed from seven labor skill categories, $K_{is}(t)$ is the sector's capital stock in period t, ρ_{is} is a parameter specific to farms of type (i, s) that is related to the elasticity of substitution between capital and the labor aggregate, γ_{is} is a parameter that is less than unity because of the absence of land from the agricultural production functions (more about this below), k is a parameter, $L_{is\lambda}(t)$ is the number of people in skill category λ who work in sector (i, s) in period t, n_{is} is the number of skill categories utilized on the (i, s) farm type, and the n_{is} exponents, $\beta_{is\lambda}$, sum to unity.

Outputs of the nine service sectors are determined by special assumptions. For the most part, output growth between periods is assumed to depend upon the level of the ratio of the service sector's current price to the average current price of commodities produced in the nonservice sectors. Labor demands are typically computed on the assumption of fixed labor–output ratios. Interindustry purchases are incorporated into the model assuming fixed input–output coefficients.

This production structure has both a number of advantages and disadvantages. The relatively large number of sectors articulated and the formal consideration of firm sizes allows us to inquire about the pattern of production in great detail. This detail brings with it, however, certain problems. The assumption that most production functions were of the Cobb–Douglas variety was probably made to economize on data, but it precludes any non-Hicks neutral technological change. It is argued in Kelley, Williamson, and Cheetham (1972) that the factor-augmenting bias in rates of technical change may be an important factor in explaining the nature of the development process. Indeed, Williamson and Lindert (personal communication) show that understanding the factor-saving bias in technical change is a crucial element in understanding inequality trends over the course of U.S. economic development. To the extent that these arguments are correct, the omission of factor-augmenting technical change from the Adelman–Robinson model reduces its ability to analyze changes in the distribution of income properly. The lack of any technological change in nine service sectors may also cause problems.

Land is omitted from the agricultural production functions in a formal sense, but the parameters γ_{is} are assumed to be less than unity to reflect diminishing returns to agricultural labor and capital alone. In essence, the land input may be considered to be subsumed in the term $A_{is}(t)$ in the production functions.

8.2 DEMAND FOR LABOR, SUPPLY OF LABOR, AND DETERMINATION OF WAGE RATES

Given output prices, factor prices, capital stocks, technical conditions, market structure, and export constraints, firms in the Adelman–Robinson model generally demand that quantity of labor services that maximizes their profits. In most cases, the derived demand functions are straightforward and so need not be described here. There are several special circumstances, however, that are useful to discuss. In the nonagricultural sectors, the smallest firms are assumed to be self-employed unskilled individuals. Therefore, these firms have no derived demand for any other laborers. In agriculture, there are assumed to be two categories of workers, family workers who must stay on a given parcel of land during the year, and other laborers who are mobile within agricultural sectors. Also, farmers on different-sized farms face different constraints on how much nonfamily labor they can hire. Given this specification, the demand for nonfamily agricultural labor also arises from the process of farmers trying to maximize their incomes. There is no demand equation for farm family workers, and consequently no equilibrium wage rate for them is determined in the model. In the service sectors, labor demands are not derived from the assumption of profit maximization, but from a set of *ad hoc* rules described above.

Labor supply to the nonagricultural sectors takes two forms. The quantity of skilled labor is considered to be fixed during the year. The quantity of low-skilled labor available to the nonagricultural sectors is assumed to vary with the wage rate according to the following specification:

$$L^s(t) = L^{*s}(t)\left[1 + \phi_s\left(\frac{W^s(t)}{W^n(t)} - 1\right)\right] \tag{8.4}$$

where $L^s(t)$ is the supply of nonagricultural labor of skill level s in year t, $L^{*s}(t)$ is the supply of nonagricultural labor of skill level s in year t under the assumption that the wage rate is $W^n(t)$, ϕ_s is an elasticity parameter specific to skill class s, $W^s(t)$ is the actual wage rate of laborers of skill class s in period t, and $W^n(t)$ is the "normal" wage rate of workers in that group in period t. The "normal" wage is defined in the model to be essentially a price index whose level is different for each skill group.

In equation (8.4) current labor supply and current wage rates are positively related. There are three possible interpretations of this association. It is possible that labor force participation rates are positively associated with real wage rates, that hours of work per individual are positively associated with wage rates, or that the rate of migration into these urban sectors from rural areas is positively related to the wage rate. Each of these three alternatives has quite different implications for the specifications in other portions of the model. The authors seem to lean toward the last interpretation, but, as we shall see below, that interpretation is difficult to square with their migration formulation.

Next, let us consider the determination of employment and wages in the nonagricultural sectors and in the agricultural sectors. In the nonagricultural nonservice sectors, wage rates are determined in a two-step procedure. First, the average wage rate for workers of a given skill level is assumed to be that wage rate that equates the aggregate demand and aggregate supply of workers of the given skill level. In the second step, the average wage is multiplied by a set of exogenous constants to compute the wage rate specific to a given industry and to a specific firm size. Further, wages in the service industries (except personal service) are also determined by multiplying the average wage by a set of exogenous constants. Thus, 78 wage rates (26 industries by 3 firm sizes) are determined from a single aggregate wage rate.

This specification seems to be seriously flawed, particularly in the context of a model that focuses upon changes in the distribution of income. On a purely technical level, that formulation seems to violate a very basic aggregation constraint: the sum of all the labor demands of the firms at the wage rates facing them should equal the aggregate demand for labor and in equilibrium the aggregate supply of labor. However, the aggregate demand for labor by firms facing the wage rates after the multiplicative adjustment described above is not, in general, equal to the aggregate demand for labor by the same firms when they all face the average wage rate. Thus, *ex post*, the aggregate supply and demand for various grades of labor are not in equilibrium. Any attempt to force them into equilibrium by modifying the firms' demands would violate the postulate of profit maximization.

On a substantive level, it seems that assuming that 78 wage rates are determined as fixed multiples of each aggregate wage rate builds into the model a substantial amount of stability in the size distribution of income. It would surely be of some interest if the robustness of the model's conclusions concerning the distribution of income could be tested in a framework in which there is more flexibility in the relative wages of individuals with the same skill levels.

In each agricultural sector, wage rates are determined so that the demand for nonfamily labor (consistent with the hiring constraints mentioned above) is equal to the exogenously determined number of nonfamily workers in that sector. No equilibrium wage is determined for family laborers.

8.3 THE TRANSLATION OF FACTOR INCOME INTO HOUSEHOLD INCOME

The Adelman–Robinson model distinguishes 15 groups of income recipients: engineers, technicians, skilled workers, apprentices, unskilled workers, white-collar workers, government workers, self-employed workers in manufacturing, self-employed workers in service occupations, capitalists, agricultural laborers, and owners of farms of four different sizes. The income distribution in each recipient group is assumed to be lognormal. The log means and roughly half the log variances are computed from the income data described above. The other log variances are determined outside the model and are assumed to be constant.

Before we continue, it should be recalled that within recipient groups for which the log variance is computed, the entire variation in income is produced by applying an exogenous set of multipliers to the average income for members of that recipient group. The number of people at each income level will vary, of course, but a substantial portion of the determinants of the log variances are built into the model in the form of the fixed multipliers.

Given survey data on the occupational distribution of workers in households where the head is in one of the fifteen recipient groups, data on the average number of workers in households in each recipient group, and the assumption that those figures remain constant over the simulation period, it is possible to compute, in a straightforward manner, the mean incomes of households where the head is in each of the recipient groups, and the numbers of households in each group. Each of these distributions is assumed to be lognormal, with the calculated mean and log variances determined in the previous step.

It is worth pausing here to digest the meaning of this last assumption. Since roughly half the log variances in the occupational income distributions are assumed to be fixed, roughly half the log variances of the household income distributions are assumed to be fixed. The other log variances are determined in good measure by the fixed multipliers discussed above. Household income distributions are combined to form the aggregate income distribution by weighting them by the proportion of households in each of the 15 categories. As we shall see below, the Adelman–Robinson model is specified so as to make substantial changes in these weights difficult to achieve. It appears, then, that the specification of the model is biased toward the conclusion that the aggregate income distribution is quite stable. It should come as no surprise, therefore, to learn that this is indeed one of the main conclusions the authors draw from their simulations.

8.4 CONSUMPTION, SAVINGS, AND INCREASES IN MONEY BALANCES

In any year, savings are computed on the assumption that average savings rates for each recipient group are constant. These average savings rates vary across recipient groups in a given year and vary over time within groups. Still in each year, the amount saved is independent of all the intragroup distributions of income and depends only on the distribution of mean income levels between groups. A preferable treatment of savings would be the use of the extended linear expenditure system (see Lluch *et al.* 1977), which makes the current savings rate depend on relative commodity prices. In addition to savings, taxes are subtracted from the mean income in each recipient group to obtain disposable income. Taxes paid by members of a recipient group do depend on the distribution of income within the group, but whether the relation between income distribution and taxation is a quantitatively significant one remains to be seen.

After the subtraction of savings and taxes from the mean income in each recipient group, consumers are assumed to allocate their remaining income to the purchase of one of the commodities or services in the model or to new money balances. The amount of disposable income spent on new money balances may be written as

$$\Delta M_h(t) = \xi_h \Delta M(t) \tag{8.5}$$

where $\Delta M_h(t)$ is the change in the holding of money balances by members of recipient group h in year t, ξ_h is a constant specific to recipient group h, and $\Delta M(t)$ is the aggregate change in money holdings for the economy as a whole in year t. The aggregate change, in turn, may be expressed as

$$\Delta M(t) = kY(t) - M(t-1) \tag{8.6}$$

where k is the average velocity of money (assumed to depend upon the inflation rate, nominal interest rates, and a time trend), $Y(t)$ is nominal GNP in year t and $M(t-1)$ is the money supply in year $t-1$.

There are two features of this approach that are especially puzzling. First, savings and increases in money holdings are determined independently. Savings are manifested neither in the purchase of durable goods nor in increases in money holdings. What form savings take is unclear. Second, changes in a group's cash balances are independent of changes in the group's income level and of the level of its cash balances. Thus, if one group's income and savings decreased, it still might increase its monetary holdings. A better specification would be one that derived each group's cash balances from information on the group's economic condition and then aggregated across groups and firms to determine aggregate money holdings.

Income available for commodity consumption, then, is obtained by subtracting from the recipient group's mean income, its mean savings, taxes, and increases in its money stocks. Consumption expenditures on goods are then determined for each recipient group from a formulation that assumes that price and income elasticities are invariant during the year. The implied system of demand equations unfortunately does not meet the "adding-up" criterion, so an *ad hoc* proportional adjustment is needed to ensure that expenditures sum to the income available for such expenditures. The income and price elasticities are readjusted every year in stage III of the model. Commodity consumption patterns, then, clearly depend on the mean of the within-group income distribution but are affected by other aspects of the distribution only to the extent that those aspects affect the group's level of taxation. A specification of the commodity composition of consumption that paid more attention to intragroup income distributions surely would have been more appropriate for this model.

8.5 INVESTMENT, GOVERNMENT EXPENDITURES, AND FOREIGN TRADE

The allocation of investment funds to sectors is done in stage I of the model and is discussed briefly below. Nominal investment is translated into the demands for the outputs of the various sectors using the current prices of those outputs and a fixed coefficients capital matrix that specifies the commodity composition of one unit of investment in each sector.

Real government expenditures in each year are specified exogenously.

Nominal expenditures on each sector are determined by multiplying the real expenditure level by an appropriate price index and then by a set of exogenously determined budget shares. The Adelman–Robinson model distinguishes five kinds of internationally traded goods: noncompetitive imports, competitive imports whose prices are domestically determined, exports whose prices are domestically determined, competitive imports whose prices are determined in the world market, and exports whose prices are determined in the world market. The specifications also take into account governmental export-promoting activities.

8.6 THE DYNAMIC EQUATIONS

Output prices, output quantities, factor prices, the price level, and the distribution of income are all determined in stage II of the model conditional on some initial conditions. These initial conditions are of two sorts. The first is essentially an updating of parameter values and changes in various stocks. These form stage III of the model. Stage I of the model describes the workings of the financial sector of the economy. The nominal levels of investment expenditures in each sector of the economy are determined there. In this section, we discuss the stage III equations. The financial sector specification will be briefly discussed in the following section.

In stage III of the Adelman–Robinson model, the productivity constants in the production functions are updated on the assumption of exogenously fixed rates of technological progress. The time profile of the interest rate for funds in the organized money market is exogenous and is updated in stage III. The exchange rate is modified in this portion of the model to take into account the last period's rate of inflation. Exports, imports, and tax rates vary over time in a predetermined manner.

In terms of the emphasis on income distribution in the Adelman–Robinson model, an important element in stage III is the representation of migration, both between rural and urban areas and between various occupational groups in the urban area. Unfortunately, this aspect of the model is discussed so briefly that it is difficult to ascertain exactly what the authors did. The natural growth rates of both the urban and the rural areas of Korea are determined exogenously. Since the urban growth rate is assumed to be somewhat higher than the rural growth rate, the model, as the authors realize, incorporates a certain amount of implicit rural–urban migration that is completely independent of their migration specification. Not only are rural and urban natural growth rates assumed to be fixed, but the natural growth rates of the various skill categories also appear to be exogenous. Rural migrants are assumed to come from agricultural laborers and owners of the two smallest sizes of farms. They are assumed to enter three urban labor groups: skilled workers, apprentices, and unskilled workers. No migrants are allowed to become self-employed urban workers. Further, we are not told in what proportions the rural migrants are allocated to each of those three urban labor groups. Once migrants arrive in the urban area and are assigned a sector, it appears that they remain in that sector for the remainder of the simulation period. This observation is modified to a minor extent, both for the migrants and for the other members of an occupational category, by the labor supply specification in equation (8.4).

The driving force behind migration is assumed to be the differential between the average incomes of people in the sending and receiving sectors of the economy. No mention is made of cost of living differentials or of any Harris–Todaro type considerations, nor is there any mention of where in the occupational income distributions the migrants come from or where they settle. The latter is particularly unfortunate for a model that focuses on questions pertaining to the distribution of income.

It can be seen that rural-urban migration is not a well-articulated phenomenon in the Adelman–Robinson model. This is also true of movement between urban occupations. The numbers of engineers, technicians, government workers, and self-employed urban workers all grow at exogenously given rates. Limited endogeneity is allowed only for skilled workers, apprentices, and unskilled workers.

Clearly, the migration specification here can be substantially improved by following the formulation in the Kelley–Williamson model discussed in Chapter 9.

8.7 THE FINANCIAL MARKET

Of all the models reviewed here, the Adelman–Robinson model provides the most detailed description of the financial side of the economy. What follows is a brief discussion of a quite detailed specification. The function of the financial market in the model is to allocate investment funds, in nominal terms, between sectors and firms. For the most part, investment demands are based on expectations of future output levels, output prices and factor prices. First, let us consider how these expectations are formed and then move on to consider how these expectations affect the allocation of investable funds.

Expectations concerning the rate of sectoral output growth are assumed to be identical across sectors and to depend on past growth rates. Each firm's expected share of its market is assumed to depend on its relative profitability in the previous period. The expected rate of output price change is assumed to be identical across sectors and is assumed to be an exogenous constant over the simulation period. Thus, expected rates of price change are not influenced by the observed rates in the recent past. Expected wage rates for the following year are assumed in the model to be the wage rates paid in the past year. Even if wages are rising steadily over time, firms will still maintain the expectation of stationary wage rates for each year into the future. The price of capital goods is assumed to grow at the same exogenous rate as prices in general.

Given these expectations, firms are assumed to demand two types of capital: working capital and fixed capital. The demand for working capital, in turn, is assumed to have two components: working capital that is required for the firm to have any positive level of output, and working capital above that minimum requirement. The demand for the first sort of working capital is proportional to the expected value of output and is independent of variations in the interest rate. The demand for the second sort of working capital depends both on the expected value of output and on the interest rate.

The demand for fixed capital on the part of manufacturing firms is the solution to the problem of maximizing profits given fixed output levels, output prices, factor prices and its initial capital stock. Certain government interventions are allowed here to encourage firms to increase their capital spending.

Service and agricultural sectors are treated differently. Service sectors are assumed to have a desired rate of growth of their capital stocks, which is allowed to vary with sector and firm size. Their demands for investment funds for fixed capital depend only on the expected price of capital goods and on the desired increase in their capital stocks. Investment in the agricultural sectors is assumed to be exogenously determined, and thus the discussion above does not apply to them.

The supply of funds for investment has five sources in the Adelman–Robinson model: retained earnings, household savings, foreign capital inflow, government savings, and the financial sector itself. Interest rates in the formal portion of the financial market are assumed to be set exogenously by the government and may differ by sector and by firm size. Firms are allowed to borrow as much as they please in the formal sector subject to a creditworthiness constraint. If they wish to borrow more than that, they can turn to the informal portion of the market, where they can borrow money at a higher interest rate. Equilibrium is reached when the interest rate in the informal sector of the financial market clears the market for investable funds.

8.8 THE ADELMAN–ROBINSON MODEL: SOME CONCLUDING THOUGHTS

The Adelman–Robinson model is truly a pioneering piece of research. It breaks new ground in a number of areas, but particularly in the field of income distribution analysis. It is unfortunate, therefore, that some of the specifications in that segment of the model are questionable. There is no doubt, however, that this work will have a substantial influence on future efforts in this field and that model builders will now be more sensitive to questions concerning the distribution of income than they have been hitherto.

9 THE KELLEY–WILLIAMSON REPRESENTATIVE DEVELOPING COUNTRY MODEL

The Kelley–Williamson (1980) model of a representative developing country (hereafter referred to as the RDC model to avoid confusion with the Kelley, Williamson, and Cheetham model) is an extension of the Kelley, Williamson, and Cheetham model described in Chapter 7. Like the KWC model and the Adelman–Robinson model, the RDC model is neoclassical in spirit, in that both output and factor prices are endogenous and simultaneously determined. The focus of the model is on the pattern of development of a representative small developing country. It has purposely been kept relatively simple in order to aid our understanding of the results that it will produce. Although the model contains several new features, its most innovative feature is its inclusion of goods that are not tradable between the urban and rural portions of the country. The existence of such goods implies that there could be cost-of-living differences between the urban and rural areas and, through this mechanism, has important implications for the pace of economic growth, migration, and the distribution of income. Let us now turn to the specification of the model.

9.1 THE PRODUCTION RELATIONS

The RDC model distinguishes eight sectors. Two sectors, manufacturing and agriculture, produce goods that are traded both internally and internationally. Their prices are determined in the world market and by the trade policy of the country. Skill-intensive services are assumed to be produced in the urban portion of the country and to be tradable within the country, but not externally. The outputs of the remaining sectors are assumed to be consumed locally. Three types of output produced in urban areas are completely nontradable: high-cost housing, low-cost housing, and labor-intensive services. Two types of output in the rural areas are completely nontradable: low-cost housing and labor-intensive services.

The production functions for manufacturing and skill-intensive services are of the two-level CES variety. They take the form

$$Q_i(t) = A_i(t)\{\xi_i\Phi_i(t)^{(\sigma_i-1)/\sigma_i} + (1-\xi_i)[z(t)\cdot L_i(t)]^{(\sigma_i-1)/\sigma_i}\}^{\sigma_i/(\sigma_i-1)} \tag{9.1}$$

$$\Phi_i(t) = \{\xi_i'[x(t)K_i(t)]^{(\sigma_i'-1)/\sigma_i'} + (1-\xi_i')[y(t)S_i(t)]^{(\sigma_i'-1)/\sigma_i'}\}^{\sigma_i'/(\sigma_i'-1)} \tag{9.2}$$

$$Z_i(t) = a_{iZ}\cdot Q_i(t) \tag{9.3}$$

$$Q_{ij}(t) = a_{ij}\cdot Q_i(t) \qquad j = 1, 2 \tag{9.4}$$

where the subscript i refers to either the manufacturing or the skill-intensive service sector; the subscript j refers to the other two remaining tradable-goods-producing sectors; $Q_i(t)$ is value-added in sector i in the period t; $A_i(t)$, $z(t)$, $x(t)$, and $y(t)$ are productivity constants; ξ_i, ξ_i', σ_i, and σ_i' are parameters of the two CES functions; $\Phi_i(t)$ is the aggregate capital variable in sector i in period t (as specified in equation 9.2); $L_i(t)$ is the quantity of unskilled labor employed in sector i in period t; $K_i(t)$ is the quantity of physical capital employed in sector i in period t; $S_i(t)$ is the quantity of skilled labor employed in sector i in period t; $Z_i(t)$ is the quantity of intermediate inputs purchased from abroad used in sector i in period t; a_{iZ} is a fixed parameter; $Q_{ij}(t)$ is the quantity of intermediate inputs purchased from domestic sector j for use in sector i in period t; and the a_{ij} comprise two fixed parameters for each sector i.

This two-level CES specification for value-added has a number of virtues. First, it can be used to investigate both the effects of biased factor-augmenting technological change and unbalanced technological progress across the various sectors of the economy. The literature has suggested the importance of both aspects of technological development and therefore it is certainly appropriate to incorporate a specification that can deal with both of them. The two-level CES formulation is consistent with the development literature in that it allows for complementarity between skilled labor and capital. It is certainly a strength of this formulation that it receives support from other work in the field.

It is somewhat unfortunate, however, that this sophisticated specification for value-added is combined with the simplest possible assumptions regarding intermediate inputs purchased domestically and intermediate inputs purchased from abroad. The constant-coefficients hypotheses manifested in equations (9.3) and (9.4) certainly simplify the model, but at a considerable cost in terms of plausibility. If the RDC model were like the Adelman–Robinson model in having a time horizon of only 9 years, then the fixed-coefficients assumptions could be acceptable. It is implausible to believe, however, that, over a 20- or 30-year simulation span, these input–output coefficients would remain unaltered. Further, this representation presumes that there can never be any input-saving technological change nor any substitution between domestically produced intermediate inputs and imported intermediate inputs. Over time, as technological progress occurs in value-added, but not in the use of intermediate goods, the cost of the latter will become an ever larger fraction of all gross output prices. Perhaps an example will help clarify one of my objections. In the face of rising oil prices, Brazil has decided to build a nuclear power plant to generate electricity and to produce gasohol as a fuel for automobiles. Neither of these substitutions is allowed given the current formulation of the production equations.

Value-added in the agricultural sectors is represented by a Cobb–Douglas production function, and there are again two fixed-coefficient intermediate inputs equations. The production relations are

$$Q_A(t) = A_A(t)\,[x(t)K_A(t)]^{\alpha}\,[z(t)L_A(t)]^{\beta}\,R(t)^{1-\alpha-\beta} \qquad (9.5)$$

$$Z_A(t) = a_{AZ}\,Q_A(t) \qquad (9.6)$$

$$Q_{Aj}(t) = a_{Aj}(t) \qquad j = 1, 2 \qquad (9.7)$$

where $Q_A(t)$ is agricultural value-added in period t; $A_A(t)$, $x(t)$, and $z(t)$ are productivity constants relevant for argiculture in period t; $K_A(t)$ is the quantity of physical capital used in agriculture in period t; $L_A(t)$ is the quantity of unskilled labor used in agriculture in period t; $R(t)$ is the quantity of land used in agriculture in period t; α and β are parameters; $Z_A(t)$ is the quantity of intermediate inputs purchased from abroad and used by agriculture in period t; a_{AZ} is a parameter; $Q_{Aj}(t)$ is the quantity of intermediate inputs purchased from domestic industry j for use in agriculture in period t; and j refers to either of the two other tradable-goods-producing sectors in the model.

There are several aspects of this specification that require comment here. First, it is not clear that the elasticity of substitution between unskilled labor and capital should be unity. My preference is not to impose that restriction on an *a priori* basis, but rather to treat agriculture and manufacturing more symmetrically. Second, the Cobb–Douglas production function for agricultural value-added implies that no skilled labor is ever used in agriculture. This assumption is very restrictive. Certainly commercial agricultural sectors in some developing countries employ quite skilled workers. Further, it is not impossible to conceive of a governmental policy aimed at increasing the skills of farmers. For this reason, it seems appropriate to allow skilled labor to enter the agricultural production function. Third, the assumption of fixed coefficients in the use of intermediate inputs separately for domestically and foreign produced goods is clearly inappropriate. Fertilizer use per unit of value-added certainly may increase over time. Also, it is possible that eventually some intermediate inputs that are currently purchased from abroad may be produced domestically. Finally, as the authors state, it would certainly be useful to disaggregate the agricultural sector, at least, into a commercial and noncommercial sector.

The output equations for the two labor-intensive service sectors are given by

$$Q_k(t) = \phi_k[z(t) \cdot L_k(t)] \tag{9.8}$$

where k refers to either of the two labor-intensive service sectors; $Q_k(t)$ is the output of sector k in year t; ϕ_k is a sector-specific constant; $z(t)$ is the productivity multiplier in period t; and $L_k(t)$ is the number of unskilled workers employed in sector k in year t.

The assumptions in this specification that capital is irrelevant to output and that there are constant returns to scale to labor alone seem to need justification. This is especially true since this sector produces low-cost housing and small-scale retail services where the values of inventories may be large relative to the values of output. Further, although it is true that the activities of members of the labor-intensive service activities may be privately profitable, it is not always clear that these activities are socially productive. Petty theft is common in urban slums, but should the "value-added" in this endeavor be added to aggregate output?

The production functions in the three housing sectors are straightforward. They are

$$Q_l(t) = \frac{H_l(t)}{\sigma_l} \tag{9.9}$$

where l refers to any of the three housing sectors, $Q_l(t)$ is the service flow from housing of type l in year t, $H_l(t)$ is the physical stock of housing of type l in year t, and σ_l is a sector-specific parameter.

9.2 STATIC LABOR DEMAND, LABOR SUPPLY, AND WAGE DETERMINATION

At any moment in time the supplies of unskilled labor in both urban and rural areas are assumed to be fixed, as is the supply of skilled labor in the urban areas. The demand for labor is obtained from the production relations on the assumptions of cost-minimizing behavior and perfectly competitive product markets. It is also assumed that the wage rates of the skilled workers in the two urban modern sectors are equalized, that the wage rates of rural unskilled workers are equal in the agricultural and service sectors, and that the wage rates of unskilled workers in the manufacturing and labor-intensive service sectors are equalized, but that these wage rates are not equal to the wage rates of unskilled workers in the traditional service sector. Instead, it is assumed that the wages of unskilled laborers in the urban modern sectors are always a fixed proportion above those of similar laborers in the urban labor-intensive service sector. Given these assumptions, three wage rates are computed that clear the three labor markets.

It is useful to return for a moment to the assumption that there is a fixed proportional wage differential between unskilled workers in the urban modern sectors and those in the urban traditional sector. It certainly appears in many developing countries that such a wage differential does indeed exist. It may be important, however, to understand the origin of the differential and whether it is likely to be constant over time. For example, in the Edmonston *et al.* (1976) model for Colombia, the wage of unskilled workers in the urban modern sector was determined by a minimum wage law and the wage of unskilled workers in the urban traditional sector was set essentially by market forces. Thus, the wage gap there is endogenous to some extent, depending, in part, on the size of past migration flows and, in part, on the demand for the output of the urban traditional sector.

9.3 SAVINGS AND THE COMMODITY COMPOSITION OF CONSUMPTION DEMAND

One important improvement in the RDC model over the KWC model is the use of the Lluch, Powell, and Williams (1977) extended linear expenditure system. The advantage of this approach is that savings flows are determined simultaneously with the commodity composition of consumption. In this framework, relative price changes, changes in disposable income, and changes in tax rates affect savings as well as the commodity composition of consumption.

Although the extended linear expenditure system is a very useful device for specifying demand structures, there are two caveats that are worth mentioning here. First, in the ELES system, in the long run as income increases, all income elasticities of demand asymptotically approach unity. This is certainly not realistic, and care must be taken when the simulation period is long that the implied income and price elasticities remain plausible. The second point is related to the first one. It is not clear that the "subsistence" quantities in the ELES system are independent of the level of income. Before this system is actually applied, it would be important to demonstrate the constancy of those "subsistence" quantities.

9.4 AGGREGATE SAVINGS AND THE COMPOSITION OF INVESTMENT DEMAND

Aggregate savings in the RDC model arises from three sources: the reinvestment of profits, household savings, and government savings. The entire flow of savings in a given period is assumed to be invested during the same period. Investment can take the form of increasing any of the three housing stocks in the model or increasing any of the three capital stocks. The financial arrangements surrounding increases in the stocks of housing and increases in the capital stocks, however, are quite different. Housing is assumed to be financed only out of the savings of those household groups that purchase the housing services. Further, demands for investable funds for housing are assumed to take priority over investment demands for the purpose of augmenting capital stocks.

The equation determining the demand for housing investment is

$$I_h(t) = \min \{S_h(t), a_h [D_h(t) - D_h(t-1)] + \delta_h H_h(t-1)\} \qquad (9.10)$$

where the subscript h refers both to housing of type h and to groups who demand housing of type h, $I_h(t)$ is the investment (in physical units) in housing of type h in year t, $S_h(t)$ is the current value of savings in year (t) by those groups who demand housing of type h, a_h is a sector-specific parameter, $D_h(t)$ is the demand (apparently measured in physical units) for housing of type h in period t, δ_h is a sector-specific depreciation parameter, and $H_h(t-1)$ is the stock of housing of type h in period $t-1$.

This approach, which separates investment in housing from investment in other capital, has two very important advantages over the competitive specifications discussed above. First, it captures an important aspect of the capital market in developing countries. Second, because of the connections between housing investment, migration, and the age structure of the population, this approach allows the investigation of the relationship between demographic and economic phenomena on a much more realistic level than do other models. The specification in equation (9.10) also has two problems. First, savings are measured in monetary units, while the second term in the brackets is measured in physical units. Thus, the equation asks for the minimum of two noncommensurate figures. The equation would be correct if the savings flow were deflated by the current cost of construction of housing of type h. Second, equation (9.10) may cause some undesirable intertemporal effects. An example should help clarify this. For simplicity, assume there is no depreciation and that a_h is equal to unity. Now assume that the demand for housing in period 1 substantially exceeds that for period 0, or, in words, that $D_h(1) - D_h(0)$ is positive and large, and that savings in that year is zero (any small number would do equally well here). In year 1, then, there is no investment in housing of type h. In year 2, let savings skyrocket so that it is no longer constraining and let $D_h(2) = D_h(1)$. The result is, plainly, that there is no investment in housing in year 2 either, even though $D_h(2)$ is substantially above $D_h(0)$ and savings is more than adequate to finance the desired housing. Clearly, some modification of equation (9.10) is in order.

Once housing demands are subtracted from the flow of savings, what remains is assumed to be invested in the three capital stocks. In the RDC model, those funds are allocated according to the following equations:

Minimize

$$|\tilde{r}_A^*(t) - \tilde{r}_M^*(t)(1-\tau_M)| + |\tilde{r}_A^*(t) - \tilde{r}_S^*(t)(1-\tau_S)| + |\tilde{r}_M^*(t)(1-\tau_M) - \tilde{r}_S^*(t)(1-\tau_S)| \qquad (9.11)$$

where

$$\tilde{r}_i^*(t) = \tilde{r}_i(t)(1-\delta_i) + \left[\frac{\partial \tilde{r}_i(t)}{\partial K_i(t)}\right] \cdot I_i(t) \qquad i = \text{A (agriculture), M (manufacturing), S (skill-intensive services)} \qquad (9.12)$$

and where $\tilde{r}_i(t)$ is the rate of return to an efficiency unit of capital in sector i in period t, τ_i is the tax rate for sector i, δ_i is the depreciation rate relevant for capital in sector i, $K_i(t)$ is the capital stock in sector i in period t, and $I_i(t)$ is the amount of investment (in physical units) in sector i in period t.

This specification embodies the notions that this segment of the capital market in the developing country is operating rather efficiently and that there is no relationship between the sector in which savings is generated and the sector in which it is invested. As the authors realize, this is certainly debatable.

There are two minor points worth mentioning about equation (9.12). First, $I_i(t)$ in equation (9.12) should be investment net of depreciation instead of gross investment. Second, net investment should be multiplied by a factor $(1 - \delta_i)$ to make both terms in the equation comparable. Finally, relative sector size is not taken into account in equation (9.11). It is possible to respecify the equation so that it is more important for the marginal rates of return for two larger sectors to be closer together than for those of a larger and a smaller sector.

9.5 FOREIGN TRADE, TAXATION, AND GOVERNMENT SPENDING

The specifications of foreign trade, taxation, and government spending in the RDC model are reasonably straightforward. There is no need to discuss all of them in detail. Instead, we shall cover here only the few cases where some possible questions arise concerning the specification.

The first point that requires mention in this context is the assumption in the model that the balance of payments is always in equilibrium. For the countries for which the RDC model is to be applicable, this assumption may not be a good one. Another formulation that allows at least transitory disequilibria may be fruitfully used here. The second point is the assumption that there are no economically relevant differences between governmentally produced services and privately produced services. This postulate certainly requires some justification. In many developing countries, governmentally controlled enterprises are often constrained to pay nonmarket clearing wages. If this phenomenon is sufficiently widespread, it may be worthwhile altering that specification.

In the portion of the model dealing with government revenues, it is assumed that tariff revenue is a constant fraction of total tax revenue. This formulation is used because the commodity composition of imports and exports cannot be determined within the structure of the model. Still, as a second best choice, this specification is not a very good one. It certainly eliminates from the model one of its interesting policy variables. Perhaps one way to improve this portion of the model is to include a separate equation for the imports of manufactured goods. This equation and the others in the model would imply a level of exports and thus allow the tariff rate to remain a policy variable.

One final point, which is relevant not only in this model but in the others reviewed here as well, is that government consumption is assumed to be an end in itself. There is never any consideration of the individuals who consume the publicly provided good. For example, in the migration decision no account is taken of the fact that governmentally provided services may be substantially greater in the urban areas than in the rural areas.

9.6 THE DYNAMIC SPECIFICATION

There are two aspects of the dynamic specification that are particularly interesting and novel: the notion of endogenous training and the migration rate specification. Let us deal with each of these briefly.

While population and labor force growth are taken to be exogenous to the model, the growth of the number of skilled laborers is taken to be endogenous. Firms are allowed to train skilled laborers when it is advantageous for them to do so. The equation used in the RDC model for the annual increase in the number of skilled workers is

$$\Delta S(t) = \epsilon_0 [(1+n)L^*(t-1) + nS(t-1)]^{\epsilon_1} [G(t-1)]^{\epsilon_2} [\text{percent wage premium}]^{\epsilon_3} \quad (9.13)$$

where $\Delta S(t)$ is the change in the stock of skilled laborers from period $t-1$ to period t; $\epsilon_0, \epsilon_1, \epsilon_2, \epsilon_3$ are parameters; n is the exogenous rate of growth of the population; $L^*(t-1)$ is the number of unskilled laborers in the two urban modern sectors in period $t-1$; $S(t-1)$ is the number of skilled workers in the economy in period $t-1$; $G(t-1)$ is governmental expenditures on noncapital items in period $t-1$; and "percent wage premium" is a complex expression for the ratio of the wages of skilled to those of unskilled workers.

This particular specification, however, seems as if it could be improved. One possibility would be to allow new skilled labor to come from two distinct sources: public education programs and private training programs. The number of skilled laborers resulting from public education programs should be explicitly linked to governmental expenditures on education programs, not to governmental expenditures on all noncapital items. The number of skilled laborers resulting from private training should be related to the effects on profits of increasing the stock of skilled laborers. That effect depends not only on the wage premium but on other features of the production function as well.

The migration portion of the RDC model is the strongest of any of the models reviewed here. Any future work in this area should undoubtedly begin with the insightful treatment of migration in the RDC model. The RDC formulation of the migration problem gets its strength from plausibly combining a number of empirically important features. Primary among these is the explicit recognition that there can be a substantial cost of living difference between the urban and the rural areas. In addition, the formulation takes into account the wage spectrum faced by new migrants and the probabilities that they will be able to obtain each of these wages. Rural–urban migration is assumed to continue in any given year until the real wage in agriculture is equal to the real expected urban wage rate.

The authors discuss the elaboration of their migration specification to include the effects of changes in the age structure of the population by utilizing the recent contribution of Rogers, Raquillet, and Castro (1978). This would undoubtedly make an already good thing even better.

9.7 THE RDC MODEL: SOME FINAL THOUGHTS

The RDC model takes what was a very good simulation specification and improves upon it. The current model is truly excellent. For a policy maker interested in economic–demographic interactions, the next step would be to begin with the RDC framework and build in more demographic structure. For example, the effect of governmental programs on education and health should be explicitly considered, as should the age structure of the population. Further, policy makers may well wish to follow the lead of Adelman and Robinson and consider the relationship between the functional distribution of income and the size distribution of household income. Whatever they wish to add, however, they can be confident that they will be off to a good start when their model is based on the RDC framework.

10 CONCLUDING COMMENTS

Five well-known second-generation economic–demographic simulation models are reviewed in this paper. None of them, in their current versions, can offer serious guidance to agricultural-policy makers. The reasons for this negative conclusion vary from case to case. In most instances, the models are of limited usefulness because agricultural policy was not the main concern of the model builders. Typically in such situations, agricultural production was not ignored, but rather its specification was simplified to the point where significant policy options were completely omitted. Those models in which the agricultural sector is sufficiently articulated to allow meaningful policy alternatives suffer from technical problems of such severity as to render what guidance they do give of questionable validity.

These economic–demographic simulation models are not totally without value for policy makers, however. In their present form, they are useful as pedogogical aids in teaching government officials about the kinds of long-run consequences their decisions could entail. Further, they provide an important step toward formalizing processes and structures the descriptions of which have hitherto been mainly discursive and the analyses of which have previously been mostly qualitative. Thus, past efforts at building economic–demographic simulation models, although they cannot be rated as successful for agricultural planning purposes, provide a useful foundation for future quantitative work.

Two third-generation simulation models are also reviewed here. Neither of them has a significant demographic component and neither can offer serious guidance to agricultural policy makers. They are useful in the context of this review for two reasons. First, they provide some improved representations of important aspects of the development process. Second, they give us a glimpse of the directions in which economic–demographic simulation models will probably be evolving in the future. For example, one evolutionary path is one being trod by Kelley and Williamson. Their latest model is of more general applicability than their earlier one. Instead of becoming involved in the intricacies of policy trade-offs in a given country, they have specified a model that is broadly applicable to a number of developing countries. The resulting model helps us to understand phenomena that are common to the development process in many countries, but the policy implications that result from the model are necessarily general ones.

Another evolutionary path is the one that Adelman and Robinson have begun to travel. This is the path toward detailed short-run models that have specific policy instruments built into them. These models need not have the breadth that the current models have, and they certainly have greater depth in the areas of particular interest. A third possible route of development of economic–demographic simulation models would combine the best features of both these two. At present there are no economic–demographic simulation models in which the trade-offs between long-run and short-run goals can be seriously studied. Such a model would certainly be useful to policy makers, who are more often judged on their ability to handle short-run crises than on their ability to solve long-term problems. Thus, now that the technology of model building is well known and widely diffused, we are likely to see a much greater variety of economic–demographic simulation models than we have seen in the past.

The history of economic–demographic simulation models has taught us a number of important lessons. Perhaps chief among them is the lesson that there is no such thing as a perfectly general model. Even with models of thousands of equations, researchers have been forced to make simplifying assumptions. Thus, the question of sorting out what is relevant and what is irrelevant to a particular problem is still important. What we have learned, then, is a lesson in modesty. There is no model for all seasons. But I must hasten to add that the blossoms in the springtime are often quite beautiful.

NOTES

1. Given a production function that is homogenous to degree one, output per worker can be written as a monotonically increasing function of capital per worker. Increasing the rate of growth of employment relative to the rate of growth of the capital stock decreases the amount of capital per worker compared to what it otherwise would have been and therefore decreases output per worker. Given a constant aggregate employment rate, the statement in the text follows immediately.

2. These models are

The FAO Model as implemented in "A Systems Simulation Approach to Integrated Population and Economic Planning with Special Emphasis on Agricultural Development and Employment: An Experimental Study of Pakistan," Food and Agriculture Organization, PA 4/1 INT/73/PO2 Working Paper Series No. 11, Rome, March 1976.

The Bachue-Philippines model as implemented in "Economic-Demographic Modelling For Development Planning: Bachue-Philippines," by G. B. Rodgers, M. J. D. Hopkins and R. Wery, International Labour Organization, Population and Employment Working Paper No. 45, Geneva, December, 1976.

The Simon Model as implemented in "Population Growth May Be Good For LDCs in the Long Run: A Richer Simulation Model," by Julian L. Simon, *Economic Development and Cultural Change*, Vol. 24, No. 2, January 1976, pp. 309–337.

The Tempo-II Model as presented in "Description of the Tempo II Budget Allocation and Human Resources Model," by William E. McFarland, James P. Bennett, and Richard A. Brown, General Electric-Tempo Working Paper GE73TMP-13, April 1973.

The Kelley, Williamson, and Cheetham Model as presented in *Dualistic Economic Development: Theory and History* by Allen C. Kelley, Jeffrey G. Williamson and Russell J. Cheetham, Chicago: University of Chicago Press, 1972.

3. These are large-scale industry, small-scale industry, capital goods industry, construction, small-scale (traditional) services, large-scale (modern) services, and government services.

4. The 1.25 figure is obtained by adding the increment in exports (assumed to be one unit) to the product of $\alpha_i(t)$ and the increment in $Z_i(t)$ (assumed to be 1.5 units).

5. The equation determining $\alpha_i(t)$ is

$$\alpha_i(t) = 1 - [1 - \alpha_i(t-1)]\,(1 - S)$$

where S is an exogenously determined policy variable.

6. In the Bachue model, current income has no effect on current consumption. The latter is determined by the past values of income. This point is discussed in more detail in section 3.2.

7. There is one exception to the statements that inputs do not affect outputs and that technical progress is irrelevant to output growth. This is in the case of traditional agriculture. It is assumed in Bachue that labor productivity in traditional agriculture increases at a predetermined but endogenous rate in each year. This assumption is maintained in all the versions of the model discussed below.

8. The maximum possible rate of growth of labor productivity in traditional agriculture is assumed to depend positively on the ratio of the prices of agricultural to nonagricultural goods.

9. Rodgers *et al.* (1976), pp. IV-17 and IV-18.

10. The determination of the $\theta_{id}(t)$ is discussed below.

11. Estimated income instead of actual income is used in this equation because no simultaneity is allowed in the Bachue model. Income is estimated using the assumption that income growth between year $t-1$ and year t at each decile level is identical to the growth that actually occurred between year $t-2$ and year $t-1$.

12. The function of the $Z^*(t)$ is described below.

13. This assumption is made on page IV.24. It is not clear, however, whether it is maintained for all time periods or just for 1965. In the text, we assumed the former.

14. The questions of the trade-off between growth and inequality can at least be addressed in the two variants of the model that allow some aggregate supply-side forces to operate. But those versions of the model are still not well suited to answer such questions. For example, it is still the case in those formulations that the income distribution has a small effect on savings, and that savings and investment have no direct links. Indeed, investment and the income distribution have practically no relation to one another. This aspect of the model requires modification if those trade-offs are to be seriously studied.

15. This statement is derived from equation 2 on page V.73 after applying the definition of a harmonic mean.

16. The equations used in determining the labor force are discussed in section 3.2 above, while those determining the number of households are discussed in section 3.4 below.

17. The $M_i(t)$ are determined from current income ratios relative to a lagged function of their historical values.

18. The correct equation is

$$W_{ri}(t) = \frac{V_{ri}(t)}{E_{ri}(t)} \cdot \frac{v_i(t)}{v_i(0)}$$

The $v_i(0)$ cannot all be set to unity without altering the input–output coefficients.

19. The number of households in each of the two areas is discussed below in section 3.4.

20. Coale, A. 1971. Age patterns of marriage. Population Studies 25(2):193–214.

21. Coale, A., and P. Demeny. 1966, Regional Model Life Tables and Stable Populations. Princeton, New Jersey: Princeton University Press.

22. Recall that in Tempo II sectoral outputs in period t are independent of any events in period t.

23. This scheme is more simplifed than the specification in Tempo II, which distinguishes students by sex.

24. This equation is derived from the equation in the footnote to page 19. I have taken

the liberty of changing the reference period for the proportion of users from period t to period $t - 1$.

25. Given the assumption of a Cobb–Douglas production function for the output of modern sector and that unskilled workers are, on the average, paid their marginal contribution to output, w is the constant appearing in equation (4.2) above.

26. The parameters 0.4 and 0.6 in equation (5.1) are the values assigned by Simon in the baseline simulations and are subject to variation in different runs. In the discussion of the Simon model that follows all the numerical parameters are of this character.

27. Either the classic problem of adding apples and oranges is ignored or the implicit assumption is made that relative prices forever remain fixed at unity.

28. It is not assumed in the FAO model that the government directly controls the allocation of investment funds. Instead, it is assumed that the government has complete indirect control over such allocations through the use of policy instruments not included in the model.

29. It is assumed for the sake of analysis here that the incremental quantities of intermediate inputs are held at zero. As was demonstrated above, the optimum amounts of these inputs are either zero or infinite. In the latter case, further efforts at maximizing output have no impact. When the incremental quantities of intermediate inputs are zero, the strategies for maximizing net and gross output are identical.

30. There are two statements made here, one concerning investment and the other concerning land conversion. Since the demonstrations of these two are essentially the same except for terminology, we shall concentrate here on sketching out the proof of only the first statement. Between any current year t and terminal year T there exists an investment strategy that will maximize agricultural output in the final year. Suppose that we take as given this optimal strategy for years $t + 1$ to T and with regard to expenditures on land conversion in year t. This can be done because it has been assumed that investment in agriculture in each year is exogenous. In this situation, the allocation of investment expenditures in period t that maximizes agricultural output in the terminal period clearly is part of the optimal strategy. In the FAO model, such a strategy involves investment in at most one form of agriculture. To see this, define $\Lambda(j, T)$ to be the amount of land in agriculture of type j in the terminal year T and define $\Delta Y(j, t, T)$ to be the increment in yield in agriculture of type j in year T due to a 1-dollar investment in that type of agriculture in current year t. Since both $\Lambda(j, T)$ and $\Delta Y(j, t, T)$ are fixed constants independent of the allocation of investment funds in period t, output in the terminal period is maximized simply by finding the single value of the index j that maximizes the product of $\Lambda(j, t)$ and $\Delta Y(j, t, T)$. If that product, by coincidence, is identical for more than one type of farming, then any distribution of investment funds between those sectors is optimal.

31. In the construction industry, the nonagricultural sector with the highest marginal product of capital, investment of 1 million dollars will bring a return in perpetuity of 911 thousand dollars per year, for a rate of return on such an investment of 91 percent per annum. Investment in the construction industry, however, is subject to special constraints, which are discussed in detail below.

32. This is based on the incremental yield coefficients in rainfed agriculture for 1965–1976. After 1976, the coefficient for small farms is assumed to fall, but the coefficient for large farms remains at its previous level.

33. Constant returns to scale are assumed here. Therefore, to determine the derived demand for construction arising from any amount of investment it is only necessary to multiply the derived demand per dollar by the number of dollars invested in the sector.

34. It is useful to recall in this context that the price of the output of the construction industry is not allowed to vary.
35. Investment in construction requires a certain amount of construction. If the sectors that lost most of the investment funds did not require much construction, then, the total amount of construction required in period $t-1$ could rise because of the reallocation of investment funds. This problem does not arise if, as in the Pakistani case, investment in the construction industry generates the least amount of construction per dollar.
36. The concept of the unemployment rate in the large-scale modern sectors is not a very clear one. This problem is discussed in more detail below, where we shall also present the FAO definition of the unemployment rate in those sectors.
37. The general fertility rate is the ratio of births to the number of women in the reproductive ages.
38. The relation between ρ_I and σ_I, the elasticity of substitution between capital and labor, is $\rho_I = 1/(1-\sigma_I)$.
39. There is neither a government nor a foreign trade sector in the KWC model.
40. The KWC model does not allow for unemployment. If unemployment were added to the model, then the wage rates adjusted for unemployment rates would have to be equalized.
41. The KWC model also includes a similar treatment of the possibility of urban-to-rural migration.

REFERENCES

Adelman, I., and S. Robinson. 1978. Income Distribution Policy in Developing Countries: A Case Study of Korea. Stanford, California: Stanford University Press.

Arthur, W. B., and G. McNicoll. 1975. Large-scale simulation models in population and development: What use to planners? Population and Development Review I(2).

Edmonston, B., W. C. Sanderson, and J. Sapoznikow. 1976. Welfare Consequences of Population Changes in Columbia: An Economic–Demographic Analysis. CREG Memorandum 207. Stanford, California: Stanford University.

Food and Agriculture Organization. 1976. A Systems Simulation Approach to Integrated Population and Economic Planning with Special Emphasis on Agricultural Development and Employment: An Experimental Study of Pakistan. PA 4/1 INT/73/PO2 Working Paper Series 11. Rome.

Lluch, C., A. A. Powell, and R. A. Williams. 1977. Patterns in Household Demand and Savings. New York: Oxford University Press.

Kelley, A. C., J. G. Williamson, and R. J. Cheetham. 1972. Dualistic Economic Development: Theory and History. Chicago: University of Chicago Press.

Kelley, A. C., and J. G. Williamson. 1980. Modeling Urbanization and Economic Growth. RR-80-22. Laxenburg, Austria: International Institute for Applied Systems Analysis.

McFarland, W. E., J. P. Bennett, and R. A. Brown. 1973. Description of the Tempo II Budget Allocation and Human Resources Model. General Electric–Tempo Working Paper GE73TMP-13. Santa Barbara, California: General Electric Company.

Rodgers, G. B., M. J. D. Hopkins, and R. Wery. 1976. Economic–Demographic Modelling for Development Planning: Bachue–Philippines. Population and Employment Working Paper 45. Geneva: International Labour Organization.

Rogers, A., R. Raquillet, and L. Castro. 1978. Model migration schedules and their applications. Environment and Planning A 10(5):475–502.

Simon, J.L. 1976. Population growth may be good for LDCs in the long run: A richer simulation model. Economic Development and Cultural Change 24(2):309–337.

ABSTRACTS OF OTHER IIASA PUBLICATIONS

Clark, W.C., D.D. Jones, and C.S. Holling, Lessons for Ecological Policy Design: A Case Study of Ecosystem Management. IIASA Research Report RR-80-2, February 1980.
Reprinted from *Ecological Modelling,* Vol. 7, 1979, pp. 1–53.

This paper explores the prospects for combining elements of the ecological and policy sciences to form a substantive and effective science of ecological policy design. This exploration is made through a case study whose specific focus is the management problem posed by competition between man and an insect (the spruce budworm, *Choristoneura fumiferana*) for utilization of coniferous forests in the Canadian Province of New Brunswick. We used this case study as a practical testing ground in which we examined the relative strengths, weaknesses, and complementarities of various aspects of the policy design process. Where existing approaches proved wanting, we sought to develop alternatives and to test them in turn. In particular, we used a combination of simulation modeling and topological approaches to analyze the space–time dynamics of this ecosystem under a variety of natural and managed conditions. Explicit consideration was given to the development of invalidation tests for establishing the limits of model credibility. An array of economic, social, and environmental indicators was generated by the model, enabling managers and policy makers to evaluate meaningfully the performance of the system under a variety of management proposals. Simplified versions of the models were constructed to accommodate several optimization procedures, including dynamic programming, which produced trial policies for a range of possible objectives. These trial policies were tested in the more complex model versions and heuristically modified in dialogue with New Brunswick's forest managers. We explored the role of utility functions for simplifying and contrasting policy performance measures, paying special attention to questions of time preferences and discounting. Finally, the study was shaped by a commitment to transfer the various models and policy design capabilities from their original academic setting to the desks and minds of the practicing managers and politicians. An array of workshops, model gaming sessions, and nontraditional communication formats was developed and tested in pursuit of this goal. This paper reports some specific management policies developed, and some general lessons for ecological policy design learned in the course of the study.

Clapham, W.B., Jr., R.F. Pestel, and H. Arnaszus, On the Scenario Approach to Simulation Modeling for Complex Policy Assessment and Design. IIASA Research Report RR-80-3, February 1980.
Reprinted from *Policy Sciences,* Vol. 11, 1979, pp. 157–177.

This paper reviews the major issues posed by scenario-based simulation modeling in the policy process, using agricultural policy as an example of a complex decision arena. Policy is seen as a process by which decision makers use the instruments under their control to approach the general goals of society. Models can help to choose instrument settings, evaluate policy options, and assess their appropriateness to a particular situation. But they cannot design policy; the interactions between policy makers and models are critical if modeling is to be useful in the policy process. Policy models must be oriented to the factors that focus and constrain judgments in the real world, as well as toward the substantive problems motivating analyses. These include the actors within the system, as well as the geographic and disciplinary contexts of the problems. Scenario-writing provides a way of ordering understanding and judgment about different phenomena to help users interact most effectively with a model and to insure that the perspectives of the model are most appropriate to the needs of the decision maker. It is an iterative and evolutionary process which can provide a great deal of insight into the assessment phase of policy design.

Beck, M.B., Model Structure Identification from Experimental Data. IIASA Research Report RR-80-4, February 1980.
Reprinted from *Theoretical Systems Ecology: Advances and Case Studies,* edited by E. Halfon, Academic Press, New York, 1979.

Methods for identifying the structure of dynamic mathematical models for water quality by reference to experimental field data are discussed. The context of the problem of model structure identification is described by briefly reviewing the steps involved in the overall process of system identification. These steps include experimental design; choice of model type; model structure identification; parameter estimation; and verification/validation. Two examples of approaches to solving the problem of model structure identification are presented. The first example is concerned with identifying the structure of a black box (input/output) model for the variations of gas production in the anaerobic digestion process of wastewater treatment. Correlation analysis is used as the principal method of solution, although it is found to have significant limitations for certain kinds of data. The second example addresses the more difficult problem of identifying the structure of an internally descriptive ("mechanistic") model form. The application of an extended Kalman-filtering algorithm to this problem is discussed in detail. The approach is illustrated with a model for phytoplankton–biochemical-oxygen-demand (BOD) interaction in a freshwater river system.

Clark, W.C., Spatial Structure Relationship in a Forest Insect System: Simulation Models and Analysis. IIASA Research Report RR-80-9, March 1980.
Reprinted from *Mitteilungen der Schweizerischen Entomologischen Gesellschaft, Bulletin de la Société Entomologique Suisse,* Vol. 52, 1979, pp. 235–257.

This paper analyzes relationships among dispersal, spatial heterogeneity, and local ecological processes in the spruce budworm (*Choristoneura fumiferana* CLEM.)–boreal forest system of eastern North America. A range of simulation and topological models are developed to reflect various hypotheses concerning those relationships. Model predictions are treated as guides to effective experimental design and efficient allocation of research priorities, rather than as ends in themselves. The analysis demonstrates the shortcomings of studies treating either dispersal or local processes alone, and argues instead for an integrated approach to spatial structure research in population ecology.

Clark, W.C., and C.S. Holling, Process Models, Equilibrium Structures, and Population Dynamics: On the Formulation and Testing of Realistic Theory in Ecology. IIASA Research Report RR-80-11, March 1980.
Reprinted from *Fortschritte der Zoologie,* Vol. 25(2/3), 1979, pp. 29–52.

This paper addresses problems in the formulation and testing of theory to relate structure and dynamic behavior in complex natural ecosystems. Detailed studies of spruce budworm–coniferous forest interactions in eastern Canada provide a background for the analysis. We argue that the mixed spatial and temporal scales, low density phenomena, and nonlinear interactions characteristic of most ecosystems severely limit traditional statistical approaches to theory building, while rendering most kinds of observational data irrelevant to theory evaluation and testing. We describe an alternative tradition: 1. Cast the theory as a set of "dynamic life tables," bound together by basic ecological process modules; apply available data and field experience to the parameterization of these modules. 2. Compute the consequences of the resulting theory under a wide range of conditions: quantitatively through numerical simulation and qualitatively through the use of topological manifolds. 3. Employ the manifolds to identify key structure- (as opposed to parameter-) dependent predictions of the theory. Compare these with observation, emphasizing behavior of the system and its theory in extreme natural or experimental situations.

Seo, F., and M. Sakawa, A Methodology for Environmental Systems Management: Dynamic Application of the Nested Lagrangian Multiplier Method. IIASA Research Report RR-80-12, March 1980.
Reprinted from *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. SMC-9, No. 12, 1979, pp. 794–805.

In this paper an alternative method for solving multiobjective optimization problems is presented. We are especially concerned with bridging a gap between procedures for obtaining the Pareto-optimal solutions and the "best compromised" preferred solution for the decision maker. First, the main concepts of the utility approach are briefly reviewed from the point of view of multiobjective systems analysis, and some shortages of this approach are examined. Second, a new method which we call the nested Lagrangian multiplier method (or NLM method) is introduced and compared with precedent devices for the utility approach. The theoretical background is also scrutinized. Third, the use of the NLM method for environmental systems management in the greater Osaka area is demonstrated, providing an example of dynamic application of this method. Finally, it is recalled that utilization of a mathematical optimization method for integrated plannings would simultaneously provide optimal solutions for allocation as well as evaluation problems, based on duality of mathematical programming. A stress is placed on the utilization of dual optimal solutions as a base of evaluation factors.

Majone, G., Policies as Theories. IIASA Research Report RR-80-17, April 1980.
Reprinted from *Omega The International Journal of Management Science,* Vol. 8, No. 2, 1980, pp. 151–162.

The received view of the scientific method, as represented for instance by logical positivism, has only historical interest for the specialists, but it is still widely, if implicitly, held by decision and policy analysts. On the other hand, recent developments in philosophy and the history of science, which stress the fallibility of theories and the social and historical character of scientific knowledge and criteria, have not yet been assimilated by analysts. This paper argues that these recent methodological developments offer important insights into many theoretical and professional problems facing students of policy making. Thus, an appreciation of the craft aspects of scientific inquiry not only clarifies the subtle relationship between theory and practice in any type of systematic analysis, but also suggests a conceptual model of the analyst's task that is quite different from the conventional decision-making paradigm. Again, Popperian and post-Popperian views of the evolution of knowledge are shown to be relevant to the evaluation of policies and to the study of their development. Particularly important in this respect is the notion, due to Lakatos, of problem shifts in competing research programs. Even the role of advocacy in policy arguments appears in a new light after we realize the importance of persuasion and propaganda in the history of scientific development. There are reasonably well-defined situations in which the use of persuasion, far from violating the analyst's code of professional behavior is not only unavoidable but also rationally justifiable.

Thomas, K., E. Swaton, M. Fishbein, and H.J. Otway, Nuclear Energy: The Accuracy of Policy Makers' Perceptions of Public Beliefs, IIASA Research Report RR-80-18, April 1980.
Also to be published in a special issue of *Behavioral Science.*

The risks associated with alternative energy systems, and public perceptions of these risks, have become important considerations in the formulation of energy policies. An earlier research memorandum (Otway and Fishbein 1977) reported a study of the attitudes and beliefs held by a sample of the Austrian public with respect to nuclear energy; an extension of the study to compare the beliefs held about five alternative energy sources has also been described (Thomas et al. 1980). The present research report analyzes the attitudes and underlying beliefs, with respect to nuclear energy, of senior Austrian civil servants in the Ministry responsible for energy matters, who were in a position to influence energy policies. It also reports on the accuracy of their perceptions of the attitudes and beliefs of those subgroups of the public sample most in favor and most against the use of nuclear energy. This report is based on work of the Joint IAEA/IIASA Risk Assessment Project, and thus it represents a collaboration between the International Atomic Energy Agency and the Energy Systems Program at the International Institute for Applied Systems Analysis.

Arthur, W.B., Why a Population Converges to Stability. IIASA Research Report RR-80-19, April 1980.

A central theorem in mathematical demography tells us that the age distribution of a closed population with unchanging fertility and mortality behavior must converge to a fixed and stable form. Proofs rely on ready-made theorems borrowed from linear algebra or from asymptotic transform theory, notably the Perron–Frobenius and the Tauberian theorems. But while these are efficient and expedient, they give little insight into the mechanism that forces the age distribution to converge. This paper proposes a simple argument for convergence. An elementary device allows us to view the birth sequence as the product of an exponential sequence and a weighted smoothing process. Smoothing progressively damps out the peaks and hollows in the initial birth sequence; thus the birth sequence gradually becomes exponential, and this forces the age distribution to assume a fixed and final form.

Williams, J., G. Krömer, and A. Gilchrist, The Impact of Waste Heat Release on Climate: Experiments with a General Circulation Model. IIASA Research Report RR-80-21, April 1980.
Reprinted from *Journal of Applied Meteorology,* Vol. 18, No. 12, 1979, pp. 1501–1511.

Experiments were made with the Meteorological Office general circulation model (GCM) to investigate the response of the simulated atmospheric circulation to the addition of large amounts of waste heat in localized areas. The concept of large-scale energy parks determined the scenarios selected for the five perturbation experiments. Waste heat totaling 150 or 300 TW was added to the sensible heat exchange between the surface and air at energy parks in the Atlantic and Pacific Oceans in four experiments. In a fifth experiment, 300 TW were added to a 10 m deep "ocean box" simulated beneath the energy parks. Forty-day averages of meteorological fields from the five waste heat experiments and from three control cases are compared. Model variability is estimated on the basis of the three control cases. The regional and hemispheric responses of the atmospheric circulation are discussed, with emphasis on the magnitude of the heating rates and 500 mb height changes. The main conclusions that can be drawn are that the model exhibits a nonlinear response to the waste heat input and that, in middle latitudes, the spatial scale of the response is large even though the heat input scale is small.

IIASA NEWS

Report on the Symposium on Modeling of Large-Scale Energy Systems, February 1980

Alan McDonald, *Energy Systems Program*

This symposium, which was jointly sponsored by IIASA and the International Federation for Automatic Control through its Systems Engineering Committee, had two objectives: "to introduce an international audience of decision makers, scientists, and representatives of industry to the scope, limitations, and current applications of energy models through an issue-oriented approach," and to stimulate "insight into and understanding of prospects, goals, and the role of future research." Thus the range of subject matter that was permitted both authors and the participants in panel discussions was broad. In the six years since the 1973 oil embargo, large-scale energy modeling has been a growing field, and practical experiences in building and applying such models have been accumulating at a rate exceeding our abilities to assimilate and communicate these various, and often isolated, experiences. Too, the more important issue in energy modeling is often not how to solve a certain problem or answer a certain question, but rather how to define what the problem is that needs solving, or how to decide what the question is that should be asked. Thus, participants were not restricted to addressing narrow questions, and the range of topics, perspectives, and opinions that found their way into the papers and discussions was extensive.

The program included an opening session in which introductory speeches were presented by Sektionschef Dr. W. Frank from the Austrian Federal Ministry for Trade, Commerce, and Industry, by Academician Professor M. Styrikovich from the USSR Academy of Sciences, and by Professor W.W. Hogan from Harvard University's John F. Kennedy School of Government. There were five topical sessions on these subjects: problems in exploring energy demand and conservation; integrated sets of models and their policy applications; problems of technology assessment, energy supply and use; questions of distribution and allocation of resources; and issues of decision making under uncertainty. There were also two panel discussions, one on "improvements in energy models to aid policy decisions" and another on "the relationship between economy, energy, capital and productivity."

Before describing the highlights of the sessions, I shall discuss some of the tensions that had troubled various participants in their experiences with large energy models. This discussion is not meant to resolve these tensions, and it would be misleading and presumptuous to suggest that they were resolved during the course of the symposium. Rather, the objective is to give some indication of the sorts of issues various participants had found most problematic and the different arguments and perspectives that evolved in conjunction with these issues.

1 The conflict between the desire to have an available, off-the-shelf model set handy for the decision maker vs. the reality that modelers are usually only needed and heeded in atypical, unprecedented situations

It would seem that, at least in part, some model builders are motivated by an image of an ideal model that would be at the service of a decision maker who could then ask such questions as, "What would happen if we decrease the tax on this commodity, or increase the funding for research in this area, or toughen the regulations on that technology?" And the model would be able to deliver predictions in understandable language quickly and with quantitative specifications of the uncertainties in the predictions. The attributes that one would associate with such a model are comprehensiveness, flexibility, and an enduring validity over time.

In reality, however, it is in situations that are not only unprecedented but also extremely difficult to anticipate that people are most likely to turn to modelers and their models. The problem is straightforward. The situations with which modelers are most familiar and which they can therefore most quickly and easily incorporate in their models are precisely the situations with which decision makers are most familiar and in which they are therefore less likely to turn to others for advice. The tension then is whether one should concentrate on building comprehensive, anticipatory models and not be distracted by transitory, short-run, perhaps over-rated fire-fighting needs, or whether one should develop problem-oriented models relevant to the real problems immediately confronting the decision maker.

2 Are decision makers informed by particular runs of particular models, or is the process by which we learn the lessons that can be learned from building and running large-scale energy models much more ambiguous?

Most discussions of policy models refer either explicitly or implicitly to an idealized, autonomous, all-powerful decision maker who is to be the beneficiary of the results. But most such discussions also point out that this idealization is obviously unrealistic, that power over and responsibility for policy decisions is usually widely shared, and that it is therefore more accurate to say that a decision somehow "gets made" than to say that some particular person or group makes the decision actively. A similar sort of false idealization often occurs in imagining how model results work their way into the policy process. It is often imagined that a particular run of a particular model is directed toward answering a particular question associated with some policy problem. But, in reality, we do not learn from models by learning specific answers to a series of discrete, specific questions. Rather, over a period, based on many runs of many models, sometimes redundant, sometimes complementary, we assimilate conclusions or new understandings. For example, it is now generally accepted that economic growth is still possible even with scarce energy resources, and for this awareness modelers deserve some credit. But one can hardly say that Model A demonstrates this conclusion or that Models B and C confirm it. Rather, it has emerged as a result generally suggested by the large collection of case analyses that has built up within the last decade. The tension here then is whether one builds a model oriented to the infamous client and his problem (to the extent that these are identifiable), or whether one views his model as contributing more importantly to the general education that is going on all the time and is emerging in some unsystematic way from the many special cases.

3 Is it possible to structure a model in an unbiased way, and does it matter?

The answer to the first part of this question is clearly "no." By definition models are simplifications, and the essence of simplifying is eliminating the superfluous. Thus, models are always biased by the modeler's perception of what is superfluous and what is essential. But there is one category of biases that is particularly troublesome: the ones introduced by the problem definition. In constructing a model the modeler must have in mind some purpose for it, some ability to provide answers or information relating to a question or a set of questions – and the questions that are assumed depend on the goals of the presumed model user. This argument leads to three problems. First, is the model user (if he is identifiable) capable of articulating his goals? Second, if he is capable, is he willing to do so? And third, if these goals are not provided by the actual or presumed model user, where does the model builder get them from?

The first and second problems are perhaps best illustrated by decision-analytic models that require the user (decision maker) to express his preferences quantitatively. Although there are now interview techniques that aspire to extract a suitable quantitative representation of the interviewee's preference structure, psychological research suggests that the inconsistencies between actual human mental processes and those necessarily assumed by the authors of the interview techniques are such as to make the interview results questionable. Moreover, it has often been pointed out that such interviews in reality tend to ignore or avoid possibly very important aspects of the interviewee's preferences, such as personal ambitions to fame, power, or money.

If it is unreasonable for the modeler to expect the user to provide him with the questions that the model should try to answer, it therefore falls to the modeler to do the best he can with a little help from his friends. This brings us to the second part of the original question: does it matter? Obviously a model designed to answer the question, "What is the appropriate contribution for nuclear power in 2000?" will yield one prediction of the energy mix in 2000, while another model designed to answer the question, "How can the penetration rate of solar energy technologies be maximized?" will yield a different prediction. Furthermore, the authors of these two imaginary models will each have to provide assumed environmental constraints, or assumed constraints on the enthusiasm or stubborness with which people respond to possible social transitions, and these constraints will necessarily incorporate, probably in an unsystematic way, further goals, both implicit and explicit, that they hold concerning the future of energy systems. Is this good or bad? Both views were expressed at the symposium. One perception is that models reveal starkly what would otherwise be implicit biases invisibly influencing supposedly objective results, and thus there is no harm in incorporating modelers' biases in their models. In fact, this incorporation allows us to understand our biases better and to see how they sometimes mislead us. The other perspective argues that all the impressive trappings of large-scale energy models, the mathematics, the data, the computer techniques, obscure – and even tend to legitimize – the hidden biases of the modeler.

Thus, there are two tensions raised by these questions. The first has to do with the relation between the model builder and the model user. Who can be expected to assume what level of responsibility for shaping the model, and how easy is it for the two to communicate? (Similar issues reappear in the next section.) The second concerns the degree to which biases of the modelers bias their models, and what the appropriate and effective modifications to models and how they are used might be.

4 How close should the relation between modelers and decision makers be?

It was agreed that the greater the institutional distance between the modeler and the decision maker, the greater the independence of the modeler. It was apparently also the consensus (at least among the modelers) that, in and of itself, more independence for the modeler is a good thing. At least two arguments supporting this position emerged during the course of the symposium. First, the more independent the modeler, the more likely he is to produce objective models free from the contaminating influences of subjective (and presumably illegitimate) biases introduced by the decision makers, who are supposedly more vulnerable to petty prejudices and perceptual weaknesses. Second, there is something to be contributed from both the perspective of the modeler and from that of the decision maker, and that therefore the distinctiveness of each of the two complementary perspectives should be preserved and encouraged, instead of being merged into some compromise (which will necessarily resemble the original position of the decision maker more closely). For example, it is perhaps to be preferred for a problem to be seen, at least initially, from both a long-term perspective (generally favored by modelers) and a short-term perspective (more usually associated with decision makers), rather than to be approached after a compromise on the perspective and through only one view.

Alternatively, modelers tend to emphasize strategies based on optimal allocations of resources, with less attention devoted to the distributional implications of these strategies. Decision makers, on the other hand, often concern themselves with distributional issues at the expense of allocational efficiency. Since our attempts to synthesize analytically the distributional and efficiency considerations have not met with success, preserving the distinction between the two may be wiser than enforcing a premature compromise.

The omnipresent problem of data availability provides a final example of the benefits of maintaining two perspectives. While it is beneficial if modelers are continually aware of the limited availability of real data, it is also beneficial for one to explore modeling possibilities unconstrained by the realities of data availability, for it is only through such an exercise that he can become aware of what additional data it would be especially helpful and useful to generate.

However, there are disadvantages arising from institutional distance. Specifically, the bigger the distance, the more difficult effective communication; large distances can easily lead to inappropriate problem definitions, unrealistic modeling assumptions, and unintelligible, untranslatable results.

In general, it was felt that, whatever the institutional distances are on the organizational chart, the more personal interactions throughout all phases of model development and use among modelers and decision makers the better. But even with this conclusion there remain some dangers. An example is the seemingly ideal situation where the authors of a large model have both direct access to the decision makers who use their model and prove to be particularly adept at translating the model results into something understandable and helpful to the decision makers. The problem is that in this situation there is little or no incentive to make the model easy for anyone else to use. Thus, when the modelers move on or when less accessible decision makers arrive, one may discover that the model evolved during the golden years turns out to be especially unintelligible to third parties. This example simply emphasizes the fact that a useful model must be reasonably easy for people other than the original authors to use and to interpret.

5 What level of detail in a model is appropriate, or should models be modularized so as to be capable of being operated at different levels of detail?

As was observed by Professor Hogan in his introductory speech on the first day of the symposium, two of the key characteristics of large-scale models are that they are difficult to build and difficult to use. They are difficult to build, not only because they are big and meant to be comprehensive, but also because, for these very reasons, one seldom starts building a large-scale model from scratch. Rather it is more likely that one starts with a set of existing submodels and sets about linking the different submodels together, perhaps adding a few new ones to fill the holes in the original set. The problems arise because the different existing submodels generally are each grounded on different assumptions and born of different data. To create from such incongruous bits and pieces one internally consistent model can often be at least as difficult as starting from scratch.

Large-scale models are difficult to use because, by virtue of their size and comprehensiveness, they seem to be at once both too complex for the particular problem that is being addressed, and not detailed enough in the area that is of principal concern to the user of the moment. Modularizing large models was suggested several times as a possible solution to this inherent difficulty. In this way the user has the freedom to use as simple or as complicated, as general or as detailed, a version of the model as he chooses. While modularization, by introducing an additional level of operational manipulation, may give rise to new tensions, as well as resolving existing ones, the modularization advocates at the symposium met no particular resistance.

THE OPENING SESSION

Sektionschef Frank opened the symposium by recounting some of the experiences that Austria, a relatively small country, has had with large-scale energy models. He mentioned two problems associated with modeling the energy economy of a smaller country. First, single decisions by important single enterprises within the system can by themselves affect the system substantially, in a way unlikely in a larger economy where the large number of enterprises allows modeling based on smooth predictions of average behavior. Second, the decisions made by other countries, often unpredictable, are important. The substantial impact of foreign decisions in the area of energy is, of course, something that we all are well aware of. The point here is that for smaller economies the problem is especially exacerbated.

Frank went on, however, to describe two models that Austria has developed. One simulates energy supply management in cases of emergency, while the other provides forecasts of the Austrian energy economy in order to check the plans of individual companies for overall consistency and sufficiency. A problem common to the experiences with both models was one that smaller countries have no monopoly on: data availability. However, Frank made the point that one should not hesitate to pursue the development of a model just because the necessary data are unlikely to be available. It is, in fact, the exploration of advanced models that contributes to defining which data ought to be gathered. He also discussed how modelers and politicians had interacted.

Academician Styrikovich dealt with the long-term global policy implications that are emerging from various analyses. The first is the necessity of making a gradual switch from oil and natural gas as the basis of our energy system to nuclear and coal. While conservation is a necessary part of any policy, to expect anything but considerable growth in demand is unrealistic. He anticipated three trends: increasing price differentials between peak-, intermediate-, and base-load electricity will provide incentives for developing better storage technologies on the supply side and better load-leveling techniques on the demand side; cost pressures are expected to induce a trend toward centralized heat generation, most likely involving cogeneration, despite problems associated with transmitting and storing heat; the transportation sector will have to shift to synthetic liquid fuels mainly derived from cheap coal.

He closed by stressing the importance of long-term thinking; because the lead times of big energy technologies are long, and because penetration times in the energy market are also long, long-term analysis and planning are necessities.

Professor Hogan reviewed the track record and status of modeling large-scale energy systems by looking at a series of problems that currently confront US policy makers and policy analysts and evaluated in each case how models have either succeeded or failed to clarify how these problems should be dealt with. The policy issues he cited include the role of price controls, the US's concern about reducing energy imports, environmental constraints, questions about investment capacities and employment effects, the relations between economic growth and energy use, the potential for conservation, how to allocate supplies in times of shortage, the depletion rates and ultimate exhaustion of nonrenewable resources, policy impacts on income distribution, the international dimensions of energy problems and world oil policy, political problems associated with decreased energy supply security, the relation between energy and inflation, the institutional constraints limiting the types of solutions offered or analyzed, and the general problems of the ventilation, validation, and accessibility of models employed in energy policy analyses. For the first six items in this list, Professor Hogan was able to identify contributions that had been made by models and modelers towards better understandings of the issues. Beyond this, however, the conclusion was that modelers, at least in the US, had much left to do. The challenge, in many ways, is to be willing to learn from past mistakes and not to forget the objective of energy models: to contribute to more informed and, therefore, better energy policies.

PROBLEMS IN EXPLORING ENERGY DEMAND AND CONSERVATION

The problems discussed in this session ranged from those that are familiar and easily defined – for example, data unavailability – to others which, though just as familiar, are more difficult to define. This latter category includes the crucial, but poorly understood, evolution and impact of cultural and political factors that ultimately play such an important role in defining the relations among energy, labor, capital, productivity, and the demand for energy services. The work described modeling on a global scale (Häfele), modeling at a national level (Danskin, Demirdache and Clayton), and modeling at a sectoral level (Bossier et al.). A variety of different modeling approaches was described: input–output, scenarios, econometric. Most of the papers reported analyses of developed countries (Belgium, Canada, the UK, and the USA); here there seemed to be a consensus that demand responses to conservation signals (price changes and subsidies were mentioned, but not direct regulation or changes in legal liability definitions, for example) are predicted to remain fairly slow. The one paper devoted solely to the developing countries cited the special analytical difficulties due not only to particularly severe data unavailability problems but also to the rapid changes occurring in the underlying economic structures of the energy systems being analyzed.

INTEGRATED SETS OF MODELS AND THEIR POLICY APPLICATIONS

While most of the papers in this session presented descriptions and results of various energy models (together with observations on the role of models and modelers in the policy processes of various countries and international organizations), several were devoted almost exclusively to less quantitative and more theoretical fundamental considerations. Danilov-Danilyan and Ryvkin from the USSR Institute for Systems Study, for example, evaluated the validity of a series of assumptions that are contained implicitly within most large-scale energy models, assumptions about resource limitations and the evolutionary patterns of future energy conversion technologies, about industrial development patterns and the nature of technology transfer between nations, about the validity and usefulness of the current world price structure as the basis for evaluating costs and benefits of future policy options, and about the validity of traditional economic indicators as descriptors of trends and as bases for policy. Although other papers devoted some space to discussing the special peculiarities of the nation or region that was being modeled and the theoretical implications these peculiarities had for developing the model appropriately, most of the time was spent on presenting and discussing the mathematics and data used by the different models, their overall structure, and results that had so far been generated. More specifically, models were discussed for Greece, the Netherlands, Japan, the USA, Czechoslovakia, the USSR, Mexico, Bulgaria, and the FRG, as well as for two multinational cases. In all cases the models were in some sense modular, usually composed of submodels dealing with, at least, energy demand, energy supply, environmental impacts, and economic and resource impacts. The modular approach was uniformly endorsed, the principal reason being that it enhanced the model's usability, and thereby popularity and credibility with users, by allowing it to be easily tailored to the question that needed answering without becoming too obscure or arcane.

PROBLEMS OF TECHNOLOGY ASSESSMENT, ENERGY SUPPLY, AND USE

One part of using models to evaluate various policy options usually involves some form of technology assessment, either for available technologies on the supply or demand side, or for technologies that will only be available in the future. In the first case, the technology assessment may lead to policy decisions to encourage or, as the case may be, to discourage the use of certain technologies by means of taxes, subsidies, regulations, legal redefinitions, and the like. In the second case, the policy tools are more likely to be applied to manipulating how effort is divided among various research and development projects.

More papers were devoted to technology assessments of energy supply than were devoted to energy use analyses. Models developed for analyzing nuclear strategies in Turkey, the French refining industry, and the future role of advanced oil technologies in Japan, as well as electricity generation systems in general were described. Only in a paper by Foell and Richter analyzing several energy/environment futures for Austria was a more detailed look at demand-side assessments reported.

The session included three papers describing models applicable to the problem of choosing research and development policies. The three included two cases where the models were used on a national level, an invited paper by Hoffman (USA) and a paper by Suzuki (Japan), as well as a model applied in an international setting, the International Energy Agency.

Finally, there were two papers describing models that went beyond technology assessment to provide a comprehensive description of policy options. The first, by Falecki and Ordęga, described a model system used to analyze energy supply and demand in Poland, and the second, by Mubayi et al., addressed the modeling of energy systems in developing countries.

QUESTIONS OF DISTRIBUTION AND ALLOCATION OF RESOURCES

The principal topic discussed in this session was, not surprisingly, prices. The invited paper by Beijdorff and Lukas from Shell International was the most concrete, in that it analyzed data describing the effect of the price increases of 1974 and 1979 on the final consumer demand for oil products. The conclusions were that the oil prices may have more effect on consumer behavior than they are often given credit for, or, put another way, the invisible hand of the market place may be more effective than is usually assumed. The paper went on to discuss some of the recent history of cost estimates for advanced technologies, and to offer some qualitative suggestions for improvements that might lead to more reliable estimates and therefore more informed decisions.

Although the second paper of the session also dealt with data, in this case a discussion of estimates and estimation techniques for world oil resources, it concentrated more on identifying weaknesses in the data than it did on offering conclusions or suggesting improvements. The third and fourth papers were more theoretical treatments of prices and pricing, discussing in the one case the divergence of optimal prices and market prices resulting in market resource allocations being inefficient, and in the other case what price an oil producer should charge in order to maximize his revenue, given the complicated interaction of pricing policies and the costs of oil alternatives. The final paper was also theoretical, presenting a generalized network-optimization code which, using a binary integer program to generate possible planning strategies, solves for an optimal regional energy-supply development plan.

ISSUES OF DECISION MAKING UNDER UNCERTAINTY

Three different applications of analytical methods, all taking particular account of uncertainty in many of the input parameters, were presented in this session. The first, an invited paper by Professor Adolf Birkhofer from the Technical University of Munich, described the results of the German Risk Study, which analyzed nuclear-power-plant safety. The conclusion drawn was that risk-analysis methods can be helpful in developing objective, quantitative criteria for power-plant design once a definition of acceptable levels of risk is available. While this is hardly a trivial step, it is argued that our experience with risk analysis so far has at least helped in beginning to understand what the issues surrounding all the difficulties really are.

A second application concerned electricity supply planning in Austria, given the significant uncertainties due to regulatory decisions, demand forecasting, supply technologies, fuel availability, and the like. The description included both models with which to understand the implications of these sources of uncertainty and some suggested characteristics of planning strategies adapted to such an uncertain environment.

The third application was also directed to supply-planning questions, in this case having to do with decisions of whether, when, and where to introduce different types of nuclear reactors into a supply system.

Of the two other papers presented in this session, one was more theoretical than those above while the other took as specific and as pragmatic a perspective as any paper in the symposium. The more theoretic of the two described a method capable of incorporating the sort of low-quality, fuzzy infomation about probabilities that is traditionally discounted in such modeling efforts. The method was illustrated by an application to the relative health risks of nuclear power. In the closing paper of the symposium, Professor Helmut Maier from the Technical University of Berlin discussed in detail the main events in the actual decision process surrounding the siting of a nuclear power plant in West Berlin. Nearly all the other papers in the symposium described models designed to assist in this or similar decisions, but one of the conclusions offered by Maier was that neither models, modelers, nor even their definition of the decision problem, played a very big role in this case. In view of the discussion that I have summarized at the beginning of this report, it is clear that this conclusion was in no sense a surprise to the participants. Nonetheless, it was an appropriate reminder that large-scale energy models are only one part of a complex and often mystifying process, and that exploiting their potential for information fully is by no means a trivial task.

PROCEEDINGS

Pergamon Press, Oxford, England, will publish the proceedings of the symposium early in 1981.

Performance and Output Measurement, Report on a Joint Meeting of EURO Public Sector and Health Working Groups, January 1980

Philip Aspden, *Human Settlements and Services*

In general, most governmental and public services have developed better methods for measuring and controlling the inputs to major programs than for measuring their performance and output. With the public demanding more value for its money, there is considerable stimulus to improve existing methods of performance and output measurement. This issue was the topic of a joint meeting at IIASA in January 1980 of two working groups of the European Association of Operational Research Societies (EURO): the Public Sector Working Group and the Health Working Group.

The aim of the meeting was to bring out common threads of thought and methods of working among the participants, and to identify specific ways in which the subject of performance and output measurement can be taken forward. As a means of achieving this aim, about half the meeting was devoted to parallel discussion syndicates (one for the Health Working Group and two for the Public Sector Working Groups). Each syndicate produced a short report on its discussion.

The discussion sessions were prefaced by papers from R.E. Levien (Director of IIASA), G. Arvidsson (Revisions Director, Swedish National Audit Bureau), I. Konya and G. Jeszensky (Hungarian Ministry of Health), and M. Lagergren (Secretariat for Future Studies, Sweden).

The papers by Levien, Arvidsson, and Lagergren approached the subject from a theoretical point of view; Konya and Jeszensky described two planning models for the Hungarian Health Care System. These illustrated the fact that where health systems are centrally planned and organized, then the setting of common standards or norms is feasible.

Levien, Arvidsson, and Lagergren discussed from their own experiences such topics as the problems of defining organizational goals, measuring these goals once defined, and producing proxy goals when the organizational goals are considered unmeasurable. Levien discussed three difficult issues the systems analyst must face when considering output and performance measurement: that it may be advantageous for the decision maker to be ambiguous about his goals, that there are likely to be misincentive effects of proxy measures, and that systems analysts tend to undervalue immeasurable goals.

The Swedish experience with performance analysis is relatively extensive. Arvidsson's paper described this experience and how work on health had influenced the form and style of the analysis. For instance, program budgeting ideas were introduced in the mid-sixties. An important feature of this approach is the attention paid to "economic rationality" and "effectiveness and efficiency." However, in the public sector, there are intrinsic values that can not be expressed in terms of economic effectiveness or rationality, which led government agencies to widen the concept of effectiveness. The current concept of effectiveness includes such factors as the employees' need for job satisfaction and a good work environment. Arvidsson also raised the question of who should carry out the performance analyses, the methodologists, the producers of the goods and services, or the consumers.

After considering performance and output measurement in the public sector in general terms, Lagergren drew on his experiences in defense and health in order to compare performance measurement in these two areas. One conclusion he drew was that there exists an underlying theoretical methodology that can be used successfully in different parts of the public sector. However, he also described the difficulties of improving public-sector decision making by applying better performance measurements.

The discussion syndicates addressed such questions as: Is it easier to measure output and performance for some types of organizations than others? In which direction should the technology for measuring outputs and performance be developed? While a variety of opinions was expressed, one area of agreement was that monitoring functions should begin at the lowest level in an organization, where the people are aware of what information is really needed. The process should then move up the organizational structure until it reaches the top. It was agreed that the monitoring processes established this way would be more useful than one imposed from above.

In his final address, R. Tomlinson, president elect of EURO, urged systems analysts to be realistic about the kinds of uses to which output measurement can be put. Unreal expectations must not be raised. He stressed that systems analysts should be concerned about giving good advice, and suggested that a way of achieving this is for the systems analyst to feed back on the results of his work. Thus, he can be sure that the performance measurement system is able to adapt to changing circumstances over time.

An informal account of the meeting is available as CP-80-7, from the Publications Department at IIASA.

BIOGRAPHIES

Martin Fishbein, USA

Martin Fishbein is Professor of Psychology and Research Professor in Communications at the University of Illinois at Urbana-Champaign. He received his Ph.D. in Psychology in 1961 from the University of California at Los Angeles, and has published several books on attitude theory, attitude measurement, and the attitude–behavior relationship.

Holding a Research Contract with the Joint IAEA/IIASA Risk Assessment Project, he contributed significantly to its research activities and coauthored a series of IIASA reports.

Ron Hinkle, USA

Ron Hinkle holds a Ph.D. in Psychology, and is a research associate of Professor M. Fishbein at the University of Illinois at Urbana-Champaign.

David Hughes, UK

David Hughes studied engineering science at Oxford University, graduating in 1971. He obtained his D.Phil. in Stochastic Control Theory from Oxford in 1974 and an M.Sc. in Statistics from London University in 1977. He has been with the Operational Research Service of the Department of Health and Social Security (UK) since 1974, working on resource allocation and regional planning.

His main interests are models of the social services, planning and decision making under conditions of uncertainty, and coordination and control. David Hughes is a Fellow of the Royal Statistical Society, and a member of the Operational Research Society and the Institute of Statisticians.

Dagmar Maurer, Austria

Dagmar Maurer joined IIASA in February 1976 to work with the Joint IAEA/IIASA Risk Assessment Project. Ms. Maurer studied anthropology and languages at the University of Vienna from 1969 to 1979 and is a Ph.D. candidate in Psychology. From 1969 to 1973 she worked as an interviewer with the Marketing Research Institute in Vienna.

Harry J. Otway, USA

Harry J. Otway is the Head of the Technology Assessment Sector at the Joint Research Centre, Commission of the European Communities, Ispra, Italy. He was Project Leader of the Joint IAEA/IIASA Risk Assessment Project from 1974 to 1978. Dr. Otway studied at the University of California where he obtained his Ph.D. in 1969.

From 1968 to 1972 he was with the Los Alamos Scientific Laboratory of the University of California at Los Angeles. Dr. Otway is a frequent Visiting Professor of Psychology and of Engineering at the University of Illinois.

His scientific interests include social implications of technological policies, and he has published extensively in journals and report series.

Warren C. Sanderson, USA

Warren C. Sanderson is a specialist in demographic economies. He received his Ph.D. in Economics from Stanford University, Stanford, California, in 1974 and is currently an assistant professor there.

His work includes papers on fertility control in nineteenth-century America, the economics of marital dissolution, a nonutilitarian economic model of fertility and female labor force participation behavior, and an economic–demographic simulation model of Colombia.

A volume on American historical demography is currently in press. He is a Visiting Research Scholar at IIASA.

David M. Simpson, South Africa

David M. Simpson was seconded, on a cost-free basis, from South Africa to the IAEA to work with the Joint IAEA/IIASA Risk Assessment Project from May 1977 to October 1978. Mr. Simpson, who holds a B.Sc. degree in Electrical Engineering from Manchester University, is currently working in the Licensing Branch, Standards Division, Atomic Energy Board, South Africa.

During his association with the Joint Project his main areas of interest were in establishing safety criteria and in cost–benefit analysis.

Bernard Spinrad, USA

Bernard Spinrad joined the Energy Systems Program at IIASA in August 1978. He is Professor of Nuclear Engineering at Oregon State University.

Professor Spinrad studied chemistry at Yale University and received his Ph.D. in Physical Chemistry in 1945. He was visiting Professor of Nuclear Engineering at the University of Illinois in 1964, and Director of the Division of Nuclear Power and Reactors at IAEA between 1967 and 1970.

Since 1972 he has been based at Oregon State University, acting as a consultant to industry, labor unions, universities, and national organizations. He was a member of the US National Research Council's Committee on Nuclear and Alternative Energy Sources (CONAES).

Professor Spinrad's main interests include reactor physics, reactor design, energy development problems, and nuclear safeguards.

Kerry Thomas, UK

Kerry Thomas is Lecturer in Social Psychology in the Facultyof Social Sciences, The Open University, Milton Keynes, England. She was a member of the Joint IAEA/IIASA Risk Assessment Project from September 1977 to September 1978.

Dr. Thomas received her Ph.D. in Biochemistry in 1966 and her M.Sc. in Social Psychology in 1968, both from the University of London. Her specific research interest is in the field of attitude formation, with particular attention to the use of attitude data in decision processes.

During her association with the Joint Project she conducted and supervised research in this field; Dr. Thomas has published extensively.

Hans Voigt, FRG

Hans Voigt joined IIASA for one year in October 1977 from the Siemens Research Center in Erlangen, FRG. While at the Institute, Dr. Voigt was involved in work on energy demand and the thermodynamics of energy utilization.

In 1978 he returned to Siemens where he is Chief Scientist with the Department of Technical Physics and is concerned with systems studies on applications of research results and with new fields of research. Dr. Voigt received his Diploma in Physics from the Humboldt University of Berlin in 1958 and his Doctorate in Physics from the Technical University of Munich in 1967.

His current research interests include technical physics, especially thermodynamics, energetics, and magnetics.

Andrzej Wierzbicki, Poland

Andrzej Wierzbicki received his Ph.D. in Automatic Control in 1964 and his D.Sc. in Mathematical Programming and Optimization in 1968, both from the Technical University of Warsaw. He has been with the Technical University of Warsaw since 1961.

At present he is leading the System and Decision Sciences Area at IIASA, where his main interests are optimization theory and applications, including multiobjective optimization and decision making, augmented Lagrange functions, and nondifferentiable optimization.

Professor Wierzbicki is a member of the High Council for Science and Education in Poland and of many other scientific councils, including that of the Institute of Systems Research, Polish Academy of Sciences. He is the author of several books and is associate editor of the journal *Automatica.*

IIASA
REPORTS

Volume 1 Number 2 April-June 1980
Completing Volume 1

Published quarterly

CONTENTS

IN FORTHCOMING ISSUES:

H.P. YOUNG, N. OKADA AND T. HASHIMOTO Cost Allocation in Water Resources Development – A Case Study of Sweden

I.V. GOUEVSKY, D.R. MAIDMENT AND W. SIKORSKI Agricultural Water Demands in Bulgaria's Silistra Region

H. FLOHN Possible Climatic Consequences of a Man-Made Global Warming