Journal of Applied Economics, Vol. VI, No. 1 (May 2003), 1-24

VECTOR AUTOREGRESSIONS, POLICY ANALYSIS, AND DIRECTED ACYCLIC GRAPHS: AN APPLICATION TO THE U.S. ECONOMY

Titus O. Awokuse[*]
University of Delaware
and
David A. Bessler[*]
Texas A&M University

Submitted May 2001; accepted March 2002

The paper considers the use of directed acyclic graphs (DAGs), and their construction from observational data with PC-algorithm TETRAD II, in providing over-identifying restrictions on the innovations from a vector autoregression. Results from Sims' 1986 model of the US economy are replicated and compared using these data-driven techniques. The directed graph results show Sims' six-variable VAR is not rich enough to provide an unambiguous ordering at usual levels of statistical significance. A significance level in the neighborhood of 30 % is required to find a clear structural ordering. Although the DAG results are in agreement with Sims' theory-based model for unemployment, differences are noted for the other five variables: income, money supply, price level, interest rates, and investment. Overall the DAG results are broadly consistent with a monetarist view with adaptive expectations and no hyperinflation.

JEL classification codes: C1, E1
Key words: vector autoregression, directed graphs, policy analysis

I. Introduction

Vector autoregressions (VARs) are widely used in empirical research because of their humility with respect to zero restrictions and assumed knowledge of the way the world actually works. Some (Cooley and Dwyer,

[*] Bessler (corresponding author): d-bessler@tamu.edu. A referee and a co-editor made substantive contributions to the paper. Any remaining errors are the authors' responsibility.

1998, Cooley and LeRoy, 1985, Leamer, 1985) have argued that, while VAR models may be useful for forecasting, they are not appropriate for policy analysis. As VARs (as usually applied) represent summaries of the correlation structure embedded in observational data (non-experimental data), they cannot be interpreted independently of a maintained structural model. In other words, for policy interpretations, the humility referred to in our opening sentence must be forgone in favor of explicit zero-type restrictions on at least some components of the VAR. In this paper we consider identifying restrictions on relationships among contemporaneous innovations.[1] The now common use of the Choleski decomposition to provide such restrictions is sometimes deemed inadequate because it imposes a just-identified contemporaneous structure that is not necessarily supported by economic theory or by the causal structure embedded in the data. Accuracy of policy inferences drawn from such analysis is therefore conditional on the validity of the maintained hypothesis of a particular just-identified structural form.

Sims (1986) and others have noted that when there is contemporaneous correlation among variables, the choice of an ordering in the Choleski decomposition may make a significant difference for interpretation of impulse responses and forecast error variance decompositions**.** As an alternative to the Choleski decomposition, some researchers (Sims, 1986; Bernanke, 1986; Blanchard and Quah, 1989; Leeper, Sims and Zha, 1996; Hess and Lee, 1999 and Kim, 2001) suggest the use of orthogonalizations that allow the researcher to impose over-identifying restrictions on the model. We follow the literature and label these models as Structural Vector Autoregressions (SVARs) as they rely on prior theory as the source of their identifying restrictions. Bernanke's approach achieves identification via the assumption that distinct, mutually orthogonal, behavioral shocks drive the model, and that lagged relationships among the variables are not restricted. The "Bernanke decomposition" relaxes

[1] More general identification restrictions could be considered on both contemporaneous innovations, as well as (subset restrictions) on lagged values of the variables in the VAR. The approach used in this paper follows that of Sims (1986) and Bernanke, where we focus on restrictions on the relationships among contemporaneous innovations. We should point out that directed graphs could be used for the more general identification problem, as well as for the restricted case considered here (see Pearl, 2000, for a discussion of identification and directed acyclic graphs).

the assumption of a just-identified structure for the VAR innovations; it requires imposing a particular causal ordering of the variables. This imposition may itself be arbitrary, as theory may not always yield a clear identifying structure.

In an often-cited paper, Sims (1986) showed that a VAR model on the U.S. economy could be used for policy analysis if appropriate identifying restrictions are imposed. He achieved identification by using two factorizations. First, he used a Choleski decomposition that imposed a just-identified structure. Second, he applied a more flexible identification method based on economic theory that relaxes the assumption of a just-identified structure for the economy.

Cooley and Dwyer (1998) argue that although VARs are attractive research tools for characterizing the dynamic relationships among variables without having to invoke economic theory restrictions, SVARs "are certainly not invariant to the identifying assumptions and may not be reliable as vehicles for identifying the relative importance of shocks." Sims' (1986) work is not exempt from this observation, as (apparently) he based his identifying constraints on subjective (non data-based) considerations. Here we investigate whether Sims' (1986) results continue to hold when a less subjective, more data-driven approach is applied to achieve an identifying interpretation of his six variable VAR on the U.S. economy. Specifically, identification is achieved by modeling the contemporaneous innovations from Sims' (1986) VAR model with directed acyclic graphs, as recently presented in Spirtes, Glymour, and Scheines (1993). These models are based on screening-off (to be explained below) characteristics present in correlations and partial correlations involving three or more variables.

The approach investigated here is one extreme, of allowing the data to provide motivation behind the over-identifying restrictions in structural VAR models. The approach is very much in the spirit of one of several uses of VARs discussed by Cooley and LeRoy (1985) and others. Cooley and LeRoy (1985, p. 288) write: "One can, of course reverse the sequence of theorizing and empirical testing. That is, econometricians can use VAR models to generate stylized facts about the causal orderings of macroeconomic variables that seem to be robust empirically. Then theorists would try to explain these patterns." This is not to say that DAGs have nothing to offer for more theoretically-based hypothesis testing with VAR models. Only that, at a minimum, understanding the “screening-off” characteristics present in a set of VAR innovations may be helpful in thinking about the mechanism that generated the data and in planning for future policy modeling with that data.

Results indicate that achieving model identification through the use of directed acyclic graphs can yield plausible and theoretically consistent impulse response functions that can be used in policy analysis. The paper is presented as follows. The next section examines a standard VAR model and the implications of the identification restrictions. We follow this with a brief introduction to directed acyclic graphs and recent algorithmic results of Spirtes, Glymour, and Scheines (1993). Sims’ (1986) policy model is then summarized and we offer a reconsideration of his model using directed acyclic graphs. A conclusion follows.

II. VAR Models and Identification

For a given vector of historical data X_t, a VAR can be expressed as:

$$\mathrm{X}_{\mathrm{t}} = \sum_{i=1}^{k} B_i \, X_{t-i} + C \, Z_t + u_t \tag{1}$$

where X_t and u_t are both (m x 1) random vectors, Z_t is a (q x 1) vector of non-stochastic (or strictly exogenous) variables, and B_i and C are appropriately dimensioned matrices of coefficients. The innovation term u_t is assumed to be white noise, where $E(u_t) = 0$, $\Sigma_u = E\,(u_t u_t')$ is an (m x m) positive definite matrix. The innovations u_t and u_s are independent for $s \neq t$. Although serially uncorrelated, contemporaneous correlation among the elements of u_t is possible. These observed innovations are mongrel, as they are combinations of more basic “structural” or driving sources of variation in the data. Following Bernanke, these driving sources of variation are themselves orthogonal and can be written as:

$$e_t = A \, u_t \tag{2}$$

Here zero restrictions on A are investigated to obtain an identified structural VAR.

Generally speaking, there are no easy counting rules for identifying *A*, but for a VAR in m variables if we leave more than m (m - 1) / 2 parameters free (to be estimated) the model is not identified. Doan (1993, pp. 8-10) suggests the following rule: if there is no combination of *i* and *j* ($i \neq j$) for which both A_{ij} and A_{ji} are nonzero, the model is identified. Usual innovation accounting procedures (impulse response, forecast error decompositions and historical decompositions) can be carried-out on the transformed VAR:

$$A\,X_t \;=\; \sum_{i=1}^{k} A\,B_i\,X_{t-i} + A\,C\,Z_t + A\,u_t \tag{3}$$

This paper's contribution is in the application of the directed acyclic graphs as an aid to identifying structural VAR models. Before discussing model specification and estimation, a brief overview of directed acyclic graphs is presented.

III. Directed Acyclic Graphs (DAGs)

Directed acyclic graphs exploit a non-time sequence asymmetry in causal relations. Consider a causally sufficient set of three variables *X*, *Y*, and *Z*. We illustrate a causal fork, *X* causes both *Y* and *Z*, as: $Y \leftarrow X \rightarrow Z$. Here the unconditional association between *Y* and *Z* is nonzero (as both *Y* and *Z* have a common cause in *X*), but the conditional association between *Y* and *Z*, given knowledge of the common cause *X*, is zero: a common cause screens-off association between its joint effects. Illustrate the inverted causal fork, both *X* and *Z* cause *Y*, as: $X \rightarrow Y \leftarrow Z$. Here the unconditional association between *X* and *Z* is zero, but the conditional association between *X* and *Z* given the common effect *Y* is not zero: a common effect does not screen-off association between its joint causes. These screening-off attributes of causal relations are captured in the literature of directed graphs.[2]

A directed graph is a picture representing the causal flow among a set of variables. More formally**,** it is an ordered triple $< V, M, E >$ where *V* is a

[2] Orcutt (1952), Simon (1953), Reichenbach (1956) and Papineau (1985) offer more detailed discussion of these screening-off asymmetries in causal relations. For a description of other causal asymmetries see Hausman (1998).

non-empty set of vertices (variables), *M* is a non-empty set of marks (symbols attached to the end of undirected edges), and *E* is a set of ordered pairs. Each member of *E* is called an edge. Vertices connected by an edge are said to be adjacent. If we have a set of vertices {*A*, *B*, *C*, *D*}: (i) the undirected graph contains only undirected edges (e.g., *A*—*B*); (ii) a directed graph contains only directed edges (e.g., $B \rightarrow C$); (iii) an inducing path graph contains both directed edges and bi-directed edges ($C \leftrightarrow D$); (iv) a partially oriented inducing path graph contains directed edges (→), bi-directed edges (↔), non-directed edges (o — o) and partially directed edges (o →). A directed acyclic graph is a directed graph that contains no directed cyclic paths (an acyclic graph contains no vertex more than once). Only acyclic graphs are used in the paper.

Directed acyclic graphs are designs for representing conditional independence as implied by the recursive product decomposition:

$$Pr\ (x_1, x_2, x_3, \ldots x_n) = \prod_{i=1}^{n} Pr\ (x_i \mid pa_i) \tag{4}$$

where *Pr* is the probability of vertices $x_1, x_2, x_3, \ldots x_n$ and pa_i the realization of some subset of the variables that precede (come before in a causal sense) X_i in order ($X_1, X_2, \ldots, X_n$). Pearl (1995) proposes d-separation as a graphical characterization of conditional independence. That is, d-separation characterizes the conditional independence relations given by equation (4). If we formulate a directed acyclic graph in which the variables corresponding to pa_i are represented as the parents (direct causes) of X_i, then the independencies implied by equation (4) can be read off the graph using the notion of d-separation (defined in Pearl, 1995):

Definition: Let *X*, *Y*, and *Z* be three disjoint subsets of vertices in a directed acyclic graph *G*, and let *p* be any path between a vertex in *X* and a vertex in *Y*, where by "path" we mean any succession of edges, regardless of their directions. *Z* is said to block *p* if there is a vertex *w* on *p* satisfying one of the following: (i) *w* has converging arrows along *p*, and neither *w* nor any of its descendants are on *Z*, or, (ii) *w* does not have converging arrows along *p*, and *w* is in *Z*. Further, *Z* is said to d-separate *X* from *Y* on graph *G*, written $(X \perp\!\!\!\perp Y \mid Z)_G$, if and only if *Z* blocks every path from a vertex in *X* to a vertex in *Y*.

Geiger, Verma, and Pearl (1990) show that there is a one-to-one correspondence between the set of conditional independencies, $(X \perp\!\!\!\perp Y \mid Z)$, implied by equation (4) and the set of triples (X, Y, Z) that satisfy the d-separation criterion in graph G. Essential for this connection is the following result: if G is a directed acyclic graph with vertex set V, A and B are in V, and H is also in V, then G linearly implies the correlation between A and B conditional on H is zero if and only if A and B are d-separated given H.

Spirtes, Glymour, and Scheines (1993) have incorporated the notion of d-separation into an algorithm (PC algorithm) for building directed acyclic graphs, using the notion of sepset (defined below).

The PC Algorithm is an ordered set of commands which begins with a general unrestricted set of relationships among variables and proceeds step-wise to remove edges between variables and to direct "causal flow." The algorithm is described in Spirtes, Glymour, and Scheines (1993, p. 117). Refinements are described as the Modified PC Algorithm (Spirtes, et al., p. 166), the Causal Inference Algorithm (p. 183), and the Fast Causal Inference Algorithm (p.188). We restrict our discussion to PC algorithm, since the basic definition of a sepset is used in all and PC Algorithm is the most basic.

Briefly, one forms a complete undirected graph G on the vertex set V. The complete undirected graph shows an undirected edge between every variable of the system (every variable in V). Edges between variables are removed sequentially based on zero correlation or partial correlation (conditional correlation). The conditioning variable(s) on removed edges between two variables is called the sepset of the variables whose edge has been removed (for vanishing zero order conditioning information the sepset is the empty set). Edges are directed by considering triples $X—Y—Z$, such that X and Y are adjacent, as are Y and Z, but X and Z are not adjacent. Edges between triples: $X—Y—Z$ are directed as: $X \rightarrow Y \leftarrow Z$, if Y is not in the sepset of X and Z. If $X \rightarrow Y$, Y and Z are adjacent, X and Z are not adjacent, and there is no arrowhead at Y, then orient $Y—Z$ as $Y \rightarrow Z$. If there is a directed path from X to Y, and an edge between X and Y, then direct $(X—Y)$ as: $\mathrm{X} \rightarrow \mathrm{Y}$.

In applications, Fisher's z is used to test whether conditional correlations are significantly different from zero. Fisher's z can be applied to test for significance from zero; where $z\,(\rho\,(i, j/\,k)\,n) = 1/2\,(n - |k| - 3)^{1/2}\,ln\,\{(|1 + \rho\,(i, j/\,k)|)\,(|1\text{-}\rho\,(i, j/\,k)|)^{-1}\}$ and n is the number of observations used to estimate the correlations, $\rho\,(i, j/\,k)$ is the population correlation between series i and j

conditional on series *k* (removing the influence of series *k* on each *i* and *j*), and |*k*| is the number of variables in *k* (that we condition on). If *i*, *j*, and *k* are normally distributed and $r\ (i, j/\ k)$ is the sample conditional correlation of *i* and *j* given *k*, then the distribution of $z\ (\rho\ (i, j/\ k)\ n)$ - $z\ (r\ (i, j/\ k)\ n)$ is standard normal.

PC Algorithm can commit type I and type II errors on both edge existence (it can fail to include an edge when it should include it and can include an edge when it should not) and edge direction (it may fail to put an arrowhead at vertex *A* when it should put it at vertex *A* and it may put an arrowhead at *A* when, in fact, it should not have put an arrowhead there). Spirtes, Glymour, and Scheines (1993) have explored several versions of PC Algorithm on simulated data with respect to errors on both edge inclusion (yes or no) and direction (arrowhead at *A* or not). They conclude that there is little chance of the algorithm including an edge that is not in the "true" model. However, there is, with small sample sizes (less than say 200 observations) considerable chance that the algorithm will omit an edge that belongs in the model. Further, arrowhead commission errors (putting an arrowhead where it does not belong) appear to be more likely than edge commission errors (putting an edge where it does not belong). Accordingly, the authors conclude: "In order for the method to converge to correct decisions with probability 1, the significance level used in making decisions should decrease as the sample size increases, and the use of higher significance levels (e.g. 0.2 at sample sizes less than 100, and 0.1 at sample sizes between 100 and 300) may improve performance at small sample sizes." (Spirtes, Glymour, and Scheines, 1993, p. 161).

Applications of directed graphs to VAR model identification are not commonplace. A similar procedure has been suggested in Swanson and Granger (1997). Their procedure considers only first order conditional correlation, and involves more subjective insight by the researcher to achieve a "structural recursive ordering." One advantage of using this method of analysis is that results based on properties of the data can be compared to a priori knowledge of a structural model suggested by economic theory or subjective intuition.

IV. Illustration Using Sims' (1986) Model

To examine the importance of using a data-determined method for achieving identification of a VAR model, we estimated Sims' (1986) six variable quarterly model of the U.S. economy using two different identification methods. One model uses the standard Sims' (1980) VAR methodology where identification is achieved via use of Choleski factorization procedure. The second model uses a modification of the Bernanke factorization where contemporaneous causal path of the model innovations is determined via use of directed graphs.

The model is estimated in log levels (except interest rates and unemployment rate, which are in levels) over the period 1948/1-1979/3. The estimation period is truncated at 1979/3 to avoid the likely need for modeling the shift in money supply behavior around 1979/4, and to allow for direct comparisons of current results with Sims' (1986). The variables in the VAR system are real GNP (*Y*), real business investment (*F*), GNP price deflator (*P*), the M1 measure of money (*M*), unemployment (*U*), and Treasury-bill rates (*R*). All measures are the same as those used in Sims (1986). Four quarterly lags on each variable and a constant term are used.

The lower triangular elements of the correlation matrix (*corr*) on innovations (errors) from the four-lag VAR, fit to 127 data points, are given as equation (5). Here we list, in lower case letters, the equation innovations for each column across the top of the matrix: *y* = innovations in income, *f* = innovations in investment, *p* = innovations in price, *m* = innovations in money, *u* = innovations in unemployment, and *r* = innovations in interest rates.

$$
\begin{array}{l}
\begin{array}{cccccc} y & f & p & m & u & r \end{array} \\
\end{array}
$$

$$
corr = \begin{bmatrix}
1.000 & & & & & \\
.518 & 1.000 & & & & \\
.004 & .002 & 1.000 & & & \\
.355 & .146 & .209 & 1.000 & & \\
-.647 & -.452 & -.194 & -.329 & 1.000 & \\
.045 & .162 & -.022 & -.039 & -.173 & 1.000
\end{bmatrix} \qquad (5)
$$

It is this matrix that drives the TETRAD II search for underlying restrictions on contemporaneous innovations.

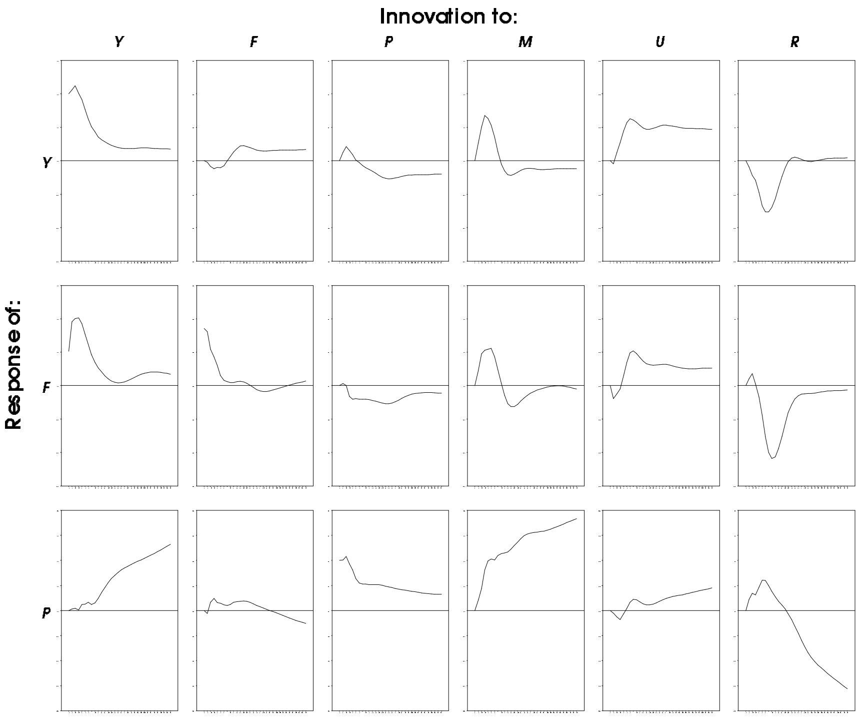
A. Model Identification Assuming a Just Identified Structure

In the first model the VAR is specified as in Sims' (1986). This allows us to replicate his impulse response functions based on a Choleski factorization (see Sims, 1986, chart 1). The variables are ordered as follows: output, investment, prices, money, unemployment rate, and interest rate. The impulse response functions obtained from this first model are presented in Figures 1.A and 1.B. Our VAR model's results and Sims' Choleski results are essentially identical. However, as we do not place instrumental priors on our VAR, responses from our Choleski decomposition will not be identical to those found in Sims (1986). Each small graph represents the response of a variable in a given row to a one-standard-deviation innovation in a variable in a given column over 32 consecutive quarters.

The dynamic effects of a (non-monetary) shock in output on real and nominal variables are presented in column 1. Positive output innovations increase output, investment, and interest rates, but decrease unemployment for about 10 quarters. An unemployment shock, column 5, is interpreted as a labor supply disturbance by Sims, capturing the complex dynamics of varying labor-force participation rate. Labor supply innovations have positive effects on output with steady increase in the first four quarters; thereafter, output remains at the higher level. While the level of unemployment rises temporarily, it returns to normal in about 8 quarters. Investment response is similar to that of output, while growth in prices is moderate. Money stock increases smoothly and remains at the higher level. The short-term interest rate is approximately constant, initially declining for a brief period then quickly returning to equilibrium levels.

Responses to money innovations are given in column 4. Real variables, income, investment, and unemployment show short-run responses, which do not persist over the long run. Money and prices show persistent long-run responses to money innovations. The delayed positive response of prices appears to be consistent with either adaptive expectations behavior or sticky prices, a point which, apparently, led Sims to suggest that commodity prices (prices set in auction markets) be added to the model to help sort-out the alternative expectations hypotheses. The weak response of real variables,

Figure 1. A. Impulse Response Functions Based on Choleski Decomposition

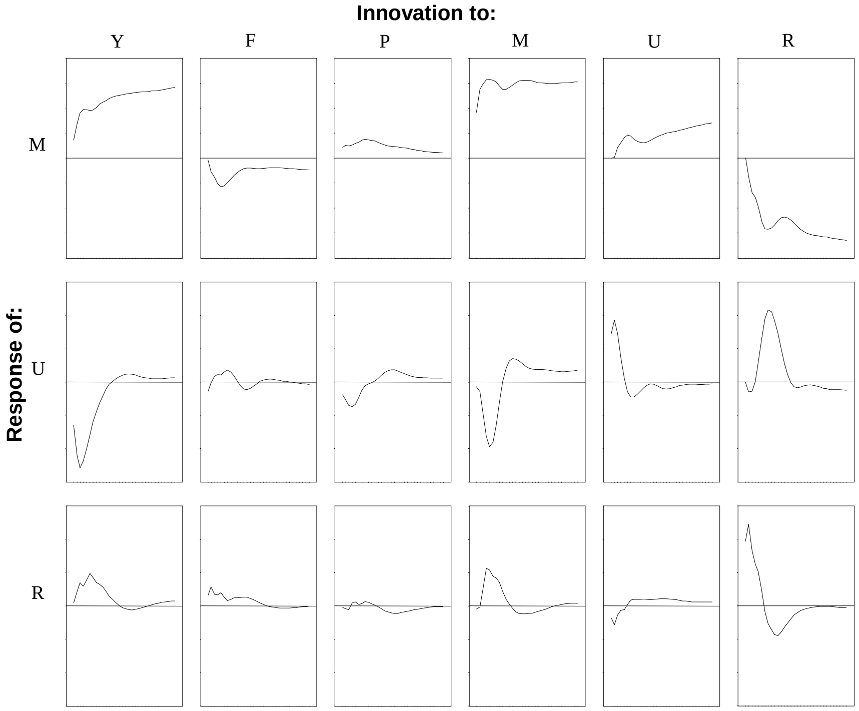


however, leads Sims (1986) to question whether these responses are consistent with a rational expectations monetarist theory. He notes: “the weakness of the real responses does not fit rational expectations monetarist theory well.”

A positive shock to interest rates yields a notable temporary decline in output, which returns to its normal level after about 12 quarters. Prices temporarily increase for about 6 quarters, and thereafter decline persistently. A strong and persistent negative response of money stock is also observed in response to innovation in interest rates. The unemployment rate momentarily declines then rises sharply for about 12 quarters before finally returning to normal.

Overall the impulse responses summarized in Figures 1.A and 1.B appear to be generally consistent with a monetarist’s view of the economy with

Figure 1. B. Impulse Response Functions Based on Choleski Decomposition



adaptive expectations (with no hyperinflation). Real variables show weak responses to money supply shocks; while prices show a persistent positive lagged response. Output responds positively and most strongly to shocks in employment.

The Choleski-generated responses are based on the contemporaneous causal ordering: innovations in output cause innovations in investment, innovations in investment cause innovations in prices, innovations in prices cause innovations in money, innovations in money cause innovations in unemployment, and innovations in unemployment cause innovations in interest rates. As an alternative to the Choleski-based responses, Sims (1986) considers theory-based interactions among innovations using the Bernanke factorization of contemporaneous correlations. Below we consider interrelations among these innovations based on directed graphs.

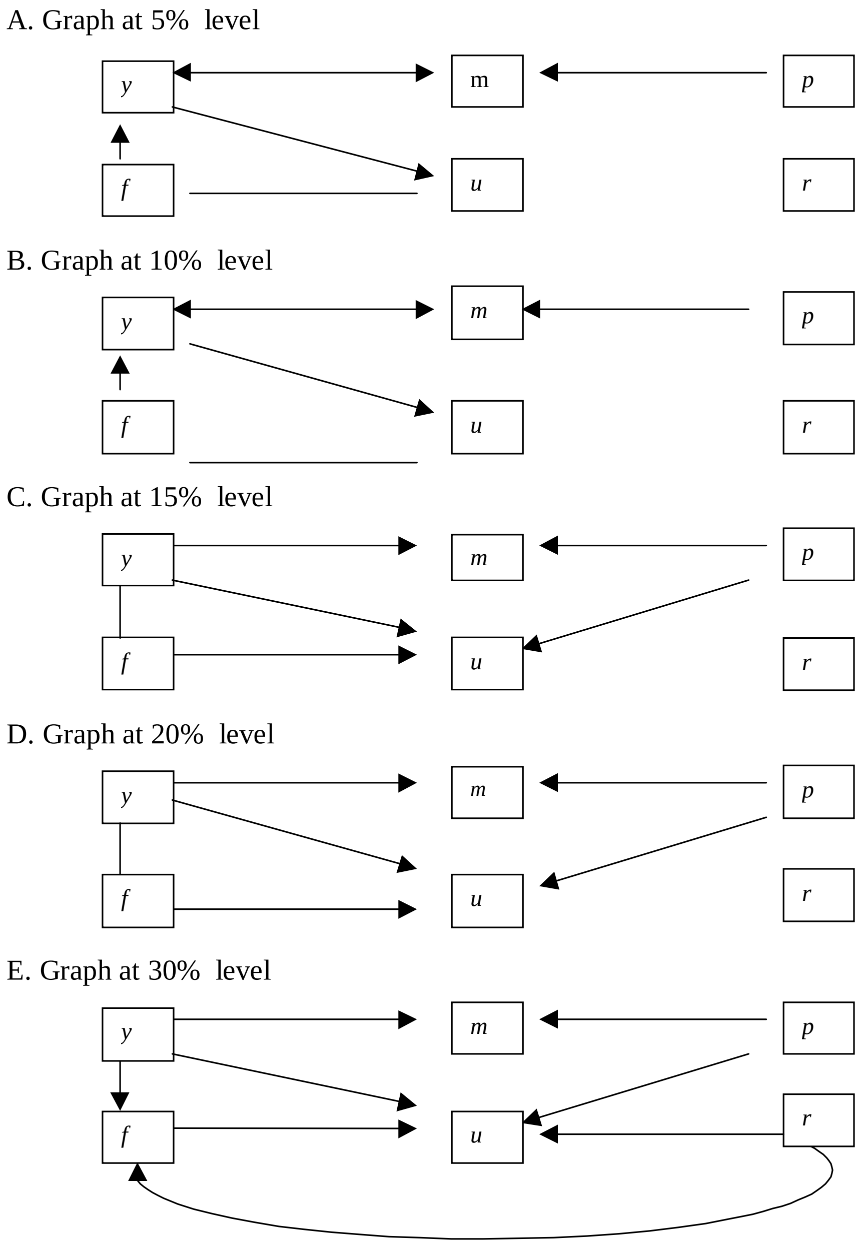
B. Directed Graph Results

The innovation correlation matrix given by equation (5) is used as the starting point for our analysis of the innovations from Sims' six-equation VAR. TETRAD II is applied to these correlations. As suggested by Spirtes, Glymour, and Scheines (1993), various levels of significance are considered in an attempt to achieve an unambiguous causal structure of the variables in contemporaneous time. Figure 2 presents graphs on innovations from Sims' (1986) six variable VAR at the following nominal levels of significance: 0.05, 0.10, 0.15, 0.20, and 0.30. As the TETRAD II search algorithm involves multiple hypothesis testing for edge removal, the final significance level is generally larger than that reported as nominal. At the 5 % and 10 % significance levels the directed edges are found as given in Panels A and B. The resulting graphs are identical, indicating directed edges from investment and money to output, and from output to money and unemployment. Directed edges are also observed running from prices to money and from money to output. However, the relationship between investment and unemployment is ambiguous, since there is an undirected edge connecting these variables (there is a relationship between investment and unemployment, but we cannot say which variable is causal).

Given the ambiguity in results at these low levels of significance, higher levels of significance of 15 % and 20 % are considered. These are given in Figure 2, Panels C and D. Although a directed edge from investment to unemployment is obtained at both of these higher levels, there is now an undirected edge between investment and output. Economic theory could be used as in Sims (1986) to direct this ambiguous causal path, but the approach will then be subject to the earlier criticism of arbitrariness. Interestingly, interest rates do not enter the system in any of the directed graphs in Panels A-D. The directed edges between prices and money, output and money, output and unemployment, and prices and unemployment seem to be stable across the 15-20 % significance levels.

Finally, as reported in Panel E, an unambiguous causal ordering is found at the 30 % level of significance. Innovations in output cause innovations in money, investment, and unemployment. Innovations in prices cause innovations in money and unemployment, while innovations from investment

Figure 2. Specification Search Using TETRAD II


A. Graph at 5% level

B. Graph at 10% level

C. Graph at 15% level

D. Graph at 20% level

E. Graph at 30% level

cause innovations in unemployment. Innovations in interest rates cause innovations in investment.

Although 30 % is a rather high significance level, it does merit discussion, as it is the lowest significance level considered which gives us an unambiguous directed graph. The alternative of using levels of say 5 % or 10 % is to conclude that the data on this six variable model are not rich enough to sort out a clear causal graph. This alternative is certainly worth considering as it is a contribution to demonstrate that Sims' (1986) six variable model does not yield a definite ordering using our directed graph techniques. However, offering the "first" unambiguous ordering in a search over alternative levels of significance allows the researcher to quantitatively assess the robustness of his/her results with respect to significance levels.

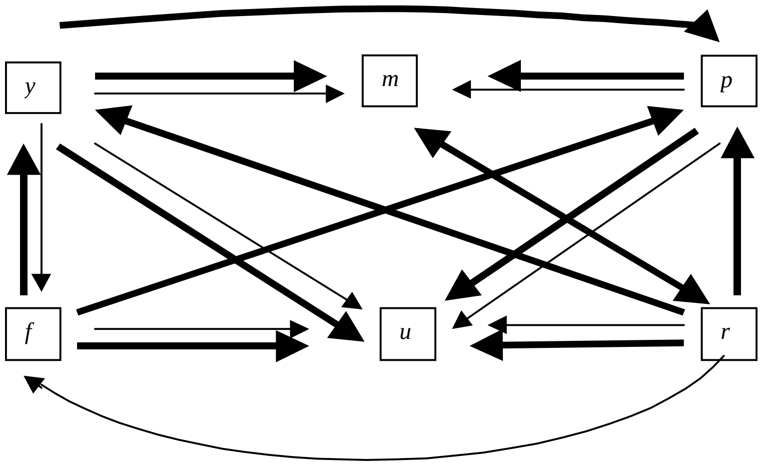
Further, Scheines et al. (1994) recommend that users of their algorithm should "vary the significance level to obtain an idea of how robust the results are. The program tends to underfit -that is, to include too few edges- at small samples. Increasing the significance level makes it easier for the program to retain edges between variables" (Scheines et al. 1994, p. 105). Given that only 127 quarterly data observations are used for this study, the suggestion to use higher significance levels is relevant in this case (although readers may suggest that our stretching their suggestion to 30 % is a priori unreasonable).

In addition to the Choleski-generated responses, Sims (1986) considers restrictions to produce theory-based impulse responses. Here he considers two models where innovations in interest rates, investment, money, prices, and output are components of the demand and supply for money. Figure 3 presents the directed graph representation of these two alternative identifications used by Sims (1986). Panel A outlines his first identification, while Panel B represents his second case. For ease of comparison, Sims' (1986) two identification scenarios, in thick bold lines, are superimposed on our model identification from Figure 2 (Panel E).

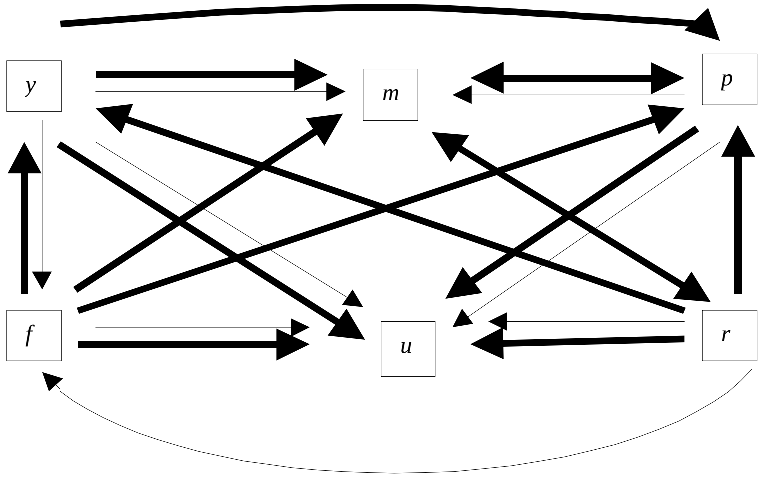
Although Sims' (1986) identification restrictions are based on economic theory and those for this study are based on data patterns, both approaches have similarities in the resulting causal structure. From Figure 3, it can be seen that both identifications allow innovations in money to respond to innovations in output and prices. The unemployment equation allows unemployment to depend on output, investment, interest rates, and prices.

**Figure 3. Sims'Identifications (Thick Lines)
and DAG Identification (Thin Lines)**

A. Sims' 1st Identification Chart and DAG at 30% level



B. Sims' 2nd Identification Chart and DAG at 30% level



However, in both panels, Sims' (1986) theory-based identifications offer several extra causal connections that seem to lack support from the data. For instance, in both identification cases, Sims (1986) suggests that innovations from interest rates cause innovations in all other variables, except investment. In contrast, the TETRAD II-based identification finds innovations in interest rates cause innovations in investment and unemployment only. Recall that a fairly high level of significance had to be used to find theses edges. Notice too that TETRAD II finds an edge running from output to investment; whereas, Sims' (1986) two alternative identifications yield the opposite causal flow; investment cause income in contemporaneous time.

Further, Sims (1986) specifies bi-directional arrows between *m* and *p* (second identification, Figure 3) and between *m* and *r* (first and second identification, Figure 3). Recall that our TETRAD II-based directed graphs too (Figure 2) resulted in bi-directional arrows (at the 5 % and 10 % levels of significance we saw *y* and *m* were bi-directed), which suggests the possibility of an omitted variable(s) or an equilibrium or feedback process.[3]

The directed edge which Sims (1986) places between innovations in interest rates (*r*) and income (*y*) does not show-up using TETRAD II, as the zero-order correlation (unconditional correlation) between innovations in interest rates and income is 0.04, with an associated p-value of 0.62 –more than double the highest-level p-value entertained in our application of TETRAD II–. Furthermore, the edge between innovations in income (*y*) and price (*p*), which Sims (1986) includes in his structural identification, does not appear in the TETRAD II model as the p-value on this edge is 0.97. In addition, Sims (1986) places edges between innovations in prices (*p*) and innovations in interest rates (*r*) and innovations in money supply (*m*) and interest rates (*r*). Zero-order correlations between these have p-values of 0.81 and 0.67, respectively, suggesting little data-generated support for these edges.

The identifying restrictions suggested by TETRAD II's graph in Figure 2, Panel E, were tested using the likelihood ratio test for over-identification as given in Doan (pp. 8-10). Given a six variable system, there are 15 lower triangular elements which can be non-zero in a just identified model, i.e.,

[3] We do not model feedback or equilibrium processes. The reader is directed to Richardson and Spirtes (1999) for a computational algorithm that can handle such cyclic graphs.

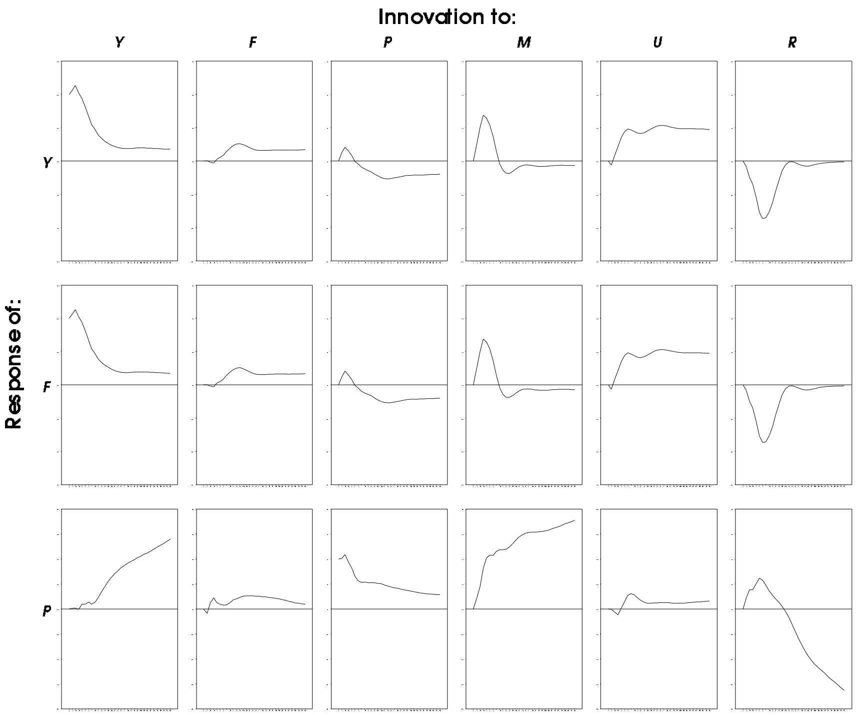
with m equal to the number of series in the VAR, we have m (m - 1) / 2 free parameters. The directed graph restrictions result in a chi-squared statistic of 2.37. With 7 degrees of freedom, we reject these zero restrictions at a p-value of 0.94, suggesting that the restrictions are consistent with the data. (We did not test Sims' (1986) ordering as it does not meet the simple Doan condition for identification).

While the above analysis suggests that several of the edges in Sims' identified model are questionable, the TETRAD II results are not without ambiguity. Probably most noticeable from Figure 2 are the reversal of causal direction as we change level of significance. At low levels of significance (0.05 and 0.10) we see that investment innovations (*f*) cause income innovations (*y*); while at the higher level of significance (0.30) we see just the opposite, innovations in investment (*f*) cause innovations in income (*y*). Furthermore, we see a bi-directed edge between innovations in income (*y*) and innovations in money (*m*) at low levels of significance; while at higher levels of significance the edge between *y* and *m* is directed as: $y \rightarrow m$. Such edge reversals are of course unsatisfying and point us in two directions. First, if we want to maintain the posture, outlined at the beginning of the paper, of relying primarily on data-based identifications, the ambiguity suggests additional data points to provide more precision on estimates of correlation and partial correlation structure. A second direction, which moves us away from our focus on databased identifications, is to rely on prior theory. Swanson and Granger (1997, p. 360) note in a discussion of their similar "structural identification procedure" that the issue of "reversibility" of causal direction among variables is "just an artifact of the contemporaneous nature of the correlation constraints that are tested." To resolve such ambiguities they suggest the use of prior knowledge based on economic theory to choose between two alternate models (1997, p. 363).

C. Innovation Accounting with the TETRAD II Suggested Structure

Figures 4.A. and 4.B. present the impulse response functions for the model identified via directed graph results. A positive shock in output (column 1) results in persistent increases in prices and money and a short-term negative response in unemployment. Comparing these responses with the responses

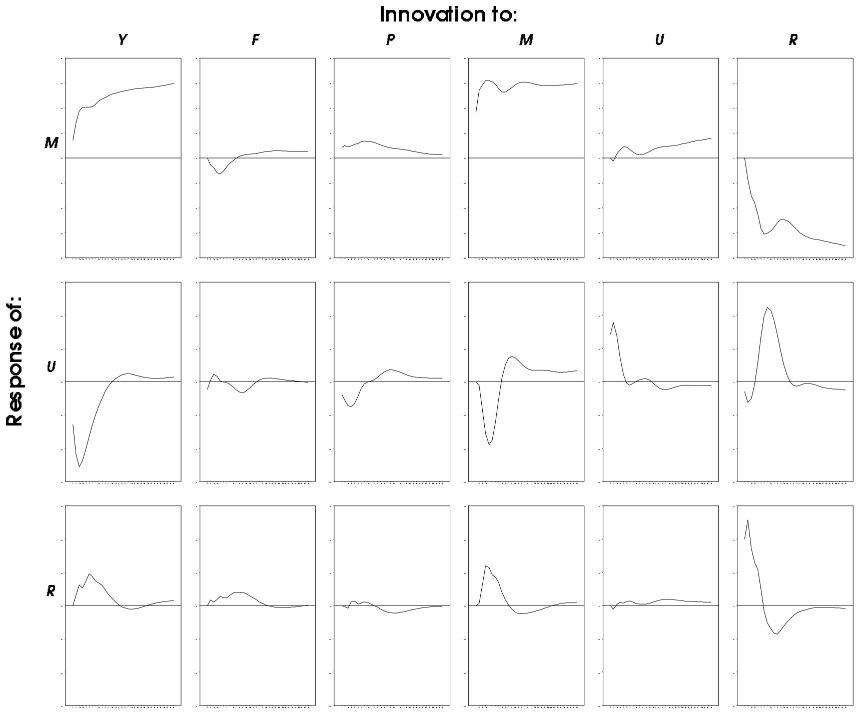
Figure 4.A. Impulse Response Functions Based on DAG at 30% Significance Level



of the same variables to innovations in investment (column 2), we see different patterns, suggesting that other components of output (consumption and government spending) may be responsible for the persistent long-run movements in prices and money and the short-term negative response of unemployment. Although, not having their measures (consumption and government spending) in this study, we cannot say more than the differences in responses are suggestive.

Positive money innovations (column 4) increase investment and output for the first 8 quarters then returning to normal within two years. Innovations in money result in sustained positive response in prices. Interest rates also respond positively in the first 3 quarters or so, thereafter returning to normal levels. Unemployment initially declines in the first year, then increases for about 6 quarters before it returns to normal levels.

Figure 4.B. Impulse Response Functions Based on DAG at 30% Significance Level



Innovations in prices and investment are not major movers of the other variables in our six variable VAR. Innovations in each of the other four variables have sustained or lasting influence on at least one other variable in our six variable system. Innovations in income have a strong and persistent impact on prices and money and considerable short-term influence on unemployment and investment. Money innovations appear to have their strongest lasting impact on prices, showing only short-term impacts (delayed by one or two periods) on the other four variables (excluding itself). Innovations in unemployment appear to be the strongest lasting influence on output. Interest rate innovations have a strong persistent influence on money and prices, both negative in the long run.

Surprisingly, the responses generated from the DAG look similar to those generated from Sims' (1986) initial Choleski factorization (Figures 1.A and

1.B). The few differences are primarily in the responses to innovations in investment. For example, consider the response of output to innovations in investment. Under the Choleski decomposition, in response to a positive shock to investment, output declines in the first eight quarters and thereafter is positive. This impulse response is (perhaps) not reasonable as we expect, a priori, that an increase in investment should result in expansion in output. Under the DAG-based decomposition, however, output responds positively to innovations in investment. This latter response is more consistent with our priors. This difference between Sims' Choleski results and our directed graph results apparently is due to differences in our respective treatments of interest rates. Sims has interest rates ordered last on the results of Figures 1.A and 1.B, while the directed graphs (Figure 2) shows interest rates as a causal factor for investment in contemporaneous time. Otherwise, the Choleski ordering used by Sims is very similar to the information flows summarized by the directed graph given in Figure 2, Panel E.

V. Concluding Remarks

The vector autoregression has found favor among many in applied econometrics for study of observational data. Among the reasons for its attractiveness is its reliance on data and avoidance of strong zero-one-type restrictions, as the VAR represents an efficient summary of the covariance patterns in historical data. However to make policy recommendations additional identifying restrictions have to be put on the VAR representation. Heretofore research workers have relied on either a Choleski factorization or theory to provide such restrictions. Both methods are subjective in the sense that the data are not given a strong role in providing explicit zero-type restrictions required for identification. This paper has asked whether results from a VAR model offered in Sims (1986) continue to hold when a less subjective, more data-driven approach, is applied to achieve an identifying interpretation of a six variable VAR on the U.S. economy.

The motivation for proceeding in this fashion was offered in the early paper by Cooley and LeRoy (1985). They suggest that one valid use of the VAR is to summarize regularities in the data, which in turn, may then motivate additional theoretical work. To date, not much work along this line has been forthcoming although Cooley and LeRoy (1985, p. 288) do cite papers by Ashenfelter and Card (1982), and Litterman and Weiss (1985) as examples of the type of research for which VAR results do generate additional theoretical work. Perhaps the reason for the lack of more studies in this vein is that both the Choleski factorization and the structural factorization involve considerable amounts of judgment on the part of the research worker. Thus it becomes problematic for analysts to know just what parts of their results are based on data and what parts based on assumed identifications. Directed acyclic graphs, while still subject to the charge of subjectivity (as we have seen here, for example, in terms of the choice of significance level), are a move in the direction in which Cooley and LeRoy point us.

Here we replicate the VAR results of an important model of Sims, where identification was achieved using a Choleski factorization. Subsequently, a second model was estimated where a contemporaneous causal ordering on the model's innovations was determined using TETRAD II's representation in terms of a directed acyclic graph. The directed graph results show Sims' six variable model not rich enough to provide an unambiguous ordering at usual levels of statistical significance. We required a significance level in the neighborhood of 30 % to find a clear structural ordering. At this rather high level of significance we found impulse response functions to be quite similar to the Choleski generated responses found by Sims (1986). These responses appear to be broadly consistent with a monetarist's view of the economy with adaptive expectations with no hyperinflation.

Additional work on type I and II errors, the possibility of multiple causal structures, and feedback and cyclic graphs is certainly warranted. Here we varied significance levels from 0.01 to 0.30 and found a number of causal patterns, one of which was a directed acyclic graph (the result found at the 30 % significance level). Questions on multiple graphs at each significance level have not been addressed. Further, we have not considered the possibility of feedback in contemporaneous time. Investigations with this algorithm (TETRAD II) and other work on cyclic graphs is now underway (see Richardson and Spirtes, 1999, for discussion of a recent algorithm for modeling cyclic graphs).

References

Ashenfelter, O., and D. Card (1982), "Time Series Representations of Economic Variables and Alternative Models of the Labour Market," *Review of Economic Studies* **49**: 761-781.

Bernanke, B.S. (1986), "Alternative Explanations of the Money-Income Correlation," *Carnegie-Rochester Conference Series on Public Policy* **25**: 49-99.

Bernanke, B.S., Gertler, M., and M. Watson (1997), "Systematic Monetary Policy and the Effects of Oil Price Shocks," *Brookings Papers on Economic Activity* **1**: 91-157.

Blanchard, O. J., and D. Quah (1989), "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review* **79**: 655-673.

Cooley, T.F., and M.D. Dwyer (1998), "Business Cycle Analysis Without Much Theory: A Look at Structural VARs," *Journal of Econometrics* **83**: 57-88.

Cooley, T.F., and S.F. LeRoy (1985), "Atheoretical Macroeconomics: A Critique," *Journal of Monetary Economics* **16**: 283-308.

Doan, T. (2000), *RATS: User's Manual*, *Version 5.0*, Evanston, Illinois, Estima.

Geiger, D., Verma, T., and J. Pearl (1990), "Identifying Independence in Bayesian Networks," *Networks* **20**: 507-533.

Hausman, D.M. (1998), *Causal Asymmetries*, New York, Cambridge University Press.

Hess, J.H., and B.S. Lee (1999), "Stock Returns and Inflation with Supply and Demand Disturbances," *The Review of Financial Studies* **12**:1203-1218.

Kim, S. (2001), "International Transmission of U.S. Monetary Shocks: Evidence from VAR's," *Journal of Monetary Economics* **48**: 339-372.

Leamer, E.E. (1985), "Vector Autoregression for Causal Inference?" *Carnegie-Rochester Conference Series on Public Policy* **22**: 225-304.

Leeper, E.M., Sims, C.A., and T. Zha (1996), "What Does Monetary Policy Do?" *Brookings Papers on Economic Activity* **2**:1-63.

Litterman, R., and L.Weiss. (1985). "Money, Real Interest Rates and Output: A Reinterpretation of Postwar U.S. Data," *Econometrica* **53**:129-56.

Orcutt, G. (1952), “Toward a Partial Redirection of Econometrics,” *Review of Economic and Statistics* **34**:195-213.

Papineau, D. (1985), “Causal Asymmetry,” *British Journal of the Philosophy of Science* **36**: 273-89.

Pearl, J. (1995), “Causal Diagrams for Empirical Research,” *Biometrika* **82**: 669-710.

Pearl, J. (2000), *Causality*, Cambridge, Cambridge University Press.

Richardson, T., and P. Spirtes (1999), “Automated Discovery of Linear Feedback Models,” in C. Glymour and G. Cooper, eds., *Computation, Causation and Discovery*, Menlo Park, AAAI Press.

Reichenbach, H. (1956), *The Direction of Time*, Berkeley, University of California Press.

Scheines, R., Spirtes, P., Glymour, C., and C. Meek (1994), *TETRAD II Tools for Causal Modeling: User’s Manual and Software*, New Jersey, Lawrence Erlbaum Associates, Inc.

Simon, H.A. (1953), “Causal Ordering and Identifiability,” in W.C. Hood and T.C. Koopmans, eds., *Studies in Econometric Method*: 49-74, New York, Wiley.

Sims, C.A (1980), “Macroeconomics and Reality,” *Econometrica* **48**: 1-48.

Sims, C.A (1986), “Are Forecasting Models Usable for Policy Analysis,” *Federal Reserve Bank of Minneapolis Quarterly Review* **10**: 2-16.

Spirtes, P., Glymour, C., and R. Scheines (1993), *Causation, Prediction, and Search*, New York, Springer-Verlag.

Swanson, N.R., and C.W.J. Granger (1997), “Impulse Response Functions Based on a Causal Approach to Residual Orthogonalization in Vector Autoregressions,” *Journal of the American Statistical Association* **92**: 357-367.

Journal of Applied Economics, Vol. VI, No. 1 (May 2003), 25-48

INTERNATIONAL TRADE, PRODUCTIVITY GROWTH, EDUCATION AND THE WAGE DIFFERENTIAL: A CASE STUDY OF TAIWAN

Hsiao-chuan Chang[*]
University of Melbourne

Submitted March 2001; accepted March 2002

The cause of changes in the wage differential between skilled and unskilled labor has been an important subject of debate for several decades. International trade and productivity growth are two main causes that have been suggested from large country studies. Recent research proposes that education is another influence. All three causes have been significantly associated with Taiwan's economic development. This paper attempts to contribute to the literature by investigating the wage differential in Taiwan, a small open economy. A Dynamic Intertemporal General Equilibrium (DIGE) model is used to perform theoretical simulation. An Error Correction Model (ECM) incorporating both short- and long-run effects is employed to accomplish the empirical examination. That education and international trade are important causes of changes in the wage differential is substantiated by Taiwanese data. Productivity growth has a significant influence on the wage differential in the short run but not in the long run.

JEL classification codes: C12, C15, C51, F10, F41, H52, J31
Key words: wage differential, dynamic intertemporal general equilibrium model, error correction model

I. Introduction

The cause of changes in the wage differential between skilled and unskilled labor has been an important issue of debate for several decades. There is a

[*] Chang: changh@unimelb.edu.au. I am greatly indebted to Warwick McKibbin and Kali Kalirajan for their valuable comments in writing this paper. I also thank Jeff Borland, Guay Lim, Peter Stemp, Alan Gunther and an anonymous referee for their helpful comments.

large amount of research that attempts to settle this issue. Essentially, the argument has been about whether international trade (reflected in trade volume, prices, measures of protection and globalization) or technological change (which induces de-industrialization and productivity growth) is the main cause. The debate is ongoing. Both labor economists (such as Katz and Autor, 1999; Wood, 1994) and trade theorists (such as Bhagwati and Dehejia, 1993; Leamer, 1994) contribute to the literature, theoretically and empirically, by proposing compelling arguments from various angles and by elaborating different methodologies. However, the more investigation one does the less conclusive the debate becomes.

Katz and Autor (1999) have developed a supply-demand-institutions (SDI) framework to assess the role of market forces (supply and demand shifts) and institutional factors in changes in the wage structure. In their discussion of market forces they analyze skill-biased technological change, globalization and de-industrialization in the determination of the wage differential. Deardorff and Hakura (1994) conducted a selective survey of the empirical literature on trade and wages, and categorized the discussion into trade volumes, prices, and measures of protection. A role for technological innovation (sometimes also referred to loosely as productivity growth) was introduced when research, for example Bound and Johnson (1992), and Lawrence and Slaughter (1993), failed to find that trade has a significant impact on wages. The majority of existing research focuses on developed or large countries and is based either theoretically on a Heckscher-Ohlin-Samuelson framework or empirically on econometric models. The econometric approaches use either reduced forms from relatively simple theoretical models or somewhat ad hoc forms, neither of which is sufficiently comprehensive.

This paper attempts to contribute to the literature by investigating the wage differential in Taiwan, a small open economy. The existing literature on the issue of trade and wages in Taiwan is quite limited. The recent work of Chen and Hsu (2001) ends with a different conclusion from both that of the studies of the U.S. and the simulation results in Chang (2000).[1] Another study, by Chan et al (1998), concludes that Taiwan's technological change is skill-biased

[1] Chen and Hsu (2001) conclude that "Taiwan's exports to less-developed countries benefit unskilled workers and thus make the income distribution more equal…".

and that progress in technology increases the wage differential. They further conclude that an increase in Taiwan's net exports increases the wage differential. In addition to providing different models compared to these existing researches, the present paper includes education investment, which is seldom discussed in the literature, as a factor underlying the wage differential.[2] An Error Correction Model (ECM) incorporating both short- and long-run effects is employed to accomplish the examination. The relevance of trade, productivity growth and education in the model as explanatory variables is substantiated by a more comprehensive theoretical model, using Dynamic Intertemporal General Equilibrium (DIGE) methodology, developed in Chang (2000). All three proposed factors have played important roles in Taiwan's economic development. This makes Taiwan an interesting case study. The results not only suggest some policy implications for the Taiwanese government but also contribute to the literature by providing comparisons with the large country cases.

This paper concludes that the results of the empirical test of the roles of trade, productivity growth and education are fairly consistent with the theoretical simulation results. That is, first, an increase in international trade increases the wage differential in both the short and the long run, with the increase being larger in the short run; second, productivity growth reduces the wage differential in the short run; and, third, an increase in government education investment decreases the wage differential in both the short and the long run, with the decrease being larger in the long run. In addition the conclusion implies that the adjustment cost of skill formation of Taiwan has been low enough to enable unskilled labor to upgrade to skilled labor and that Taiwan's productivity growth is not skill-biased.[3] It also indicates the importance of examining the cost of skill adjustment in future researches on the wage differential.

Section II illustrates a profile of the wage differential in Taiwan. Section

[2] Chan et al (1998) only use education as a criterion to separate labor into skilled and unskilled groups. In their regression, there is no explicit variable for education.

[3] This finding is opposite to the finding of Chan et al (1998).

III briefly sets out the theoretical model, Section IV demonstrates the empirical test and Section V summarises the conclusions.

II. The Wage Differential in Taiwan

This section describes the wage differential between different skill groups in Taiwan from 1978 to 1999. The monthly average wage data by education attainment are from the Manpower Utilization Survey. The real wage is the nominal wage deflated by the GDP deflator measured on the base of the price level in 1996. Conventionally, education attainment is the principal determinant of skilled and unskilled labor. Practically, the education level used to split labor into skilled and unskilled groups might affect the consequences. This section presents two versions of the split.[4] First, individuals with a degree from college or above are designated as skilled labor while the remainder are designated as unskilled labor. Second, individuals with a degree from junior college or above are designated as skilled labor while the remainder are designated as unskilled labor. To manipulate the raw data, which include several categories in the group of unskilled labor, the weighted average monthly wage is used, with the weights being the population proportion of each category in the group.[5]

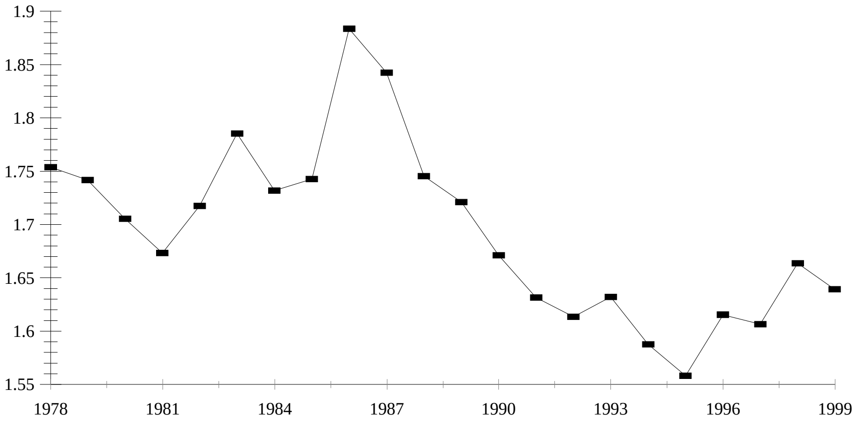
A. Degree from College or Above Designated to Skilled Labor

Figure 1 illustrates the variations in the ratio of the wage of skilled labor to the wage of unskilled labor from 1978 to 1999. From 1981 to 1986, the wage ratio shows fluctuations around an upward trend. From 1987 to 1995, the ratio follows a straight downward trend (except for 1993). This indicates that the unskilled wage grew more than the skilled wage and reflects a continuing

[4] The reason being, in Taiwan, there is a category called “junior college.” Students spend two, three, or five years to get a degree which is of a lower rank than a four-year college degree. The subjects offered in junior colleges are similar to those offered in colleges. Hence, it is considered reasonable to take a look at the case where this category is included in skilled labor. Katz and Murphy (1992) created a measurement of college and high school equivalents, which might not be a good alternative given the limited time series data.

[5] The categories are: illiterate, self-educated, primary school, junior high (including junior vocational) school, senior high school, vocational school, and/or junior college.

Figure 1. Wage Ratio: Skilled to Unskilled Labor, College or Above as Skilled Labor



shortage of unskilled labor in Taiwan. During this period, government policy on importing foreign labor becomes important.[6] After 1995, the wage ratio shows a tendency to increase.

B. Degree from Junior College or Above Designated to Skilled Labor

In this case both the growth rate and the wage ratio are similar to those in Section II.A except that the wage ratio is smaller than when junior college is not designated to skilled labor. Figure 2 illustrates the wage ratio when junior college is designated as skilled labor. Obviously the smaller wage ratio is the result of the use of a weighted average method and the fact that the wage rate resulting from junior college education is less than that resulting from college education.

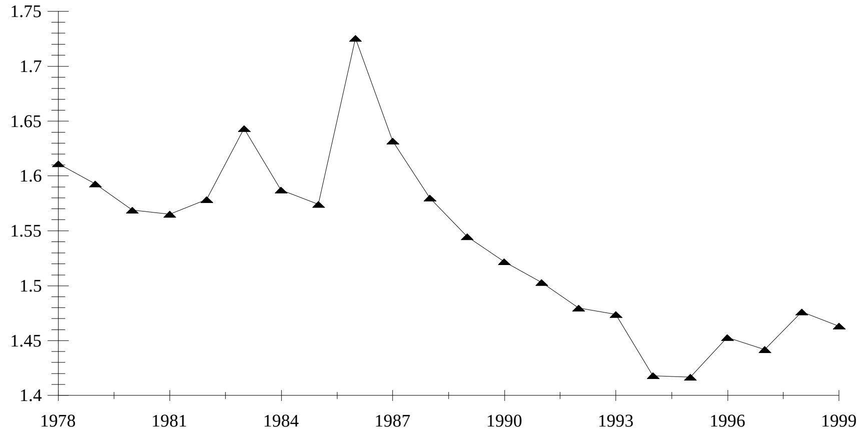
In summary, both versions show that there is no convincing evidence of a growing wage differential in Taiwan over the past two decades.

III. Theoretical Model

The conventional Heckscher-Ohlin-Samuelson model is usually used to

[6] Refer to Tsay (1995) for a detailed discussion.

Figure 2. Wage Ratio: Skilled to Unskilled Labor, Junior College or Above as Skilled Labor



investigate issues involving international trade. Among the limitations of the Heckscher-Ohlin-Samuelson model are that the quantities of factors of production are assumed fixed and that predictions of the variations in the wage differential are limited to the long-run variations. Under its simplified framework, due to the fixed quantities of production factors, it predicts that an increase in international trade decreases a country's unskilled wage and increases its skilled wage if the country exports skilled-labor intensive goods and imports unskilled-labor intensive goods. Technological progress, if it is biased toward the demand for skilled labor, leads to an increase in the wage differential. The converse also applies. The assumption of fixed endowments precludes consideration of the important role of changes in factor supplies. Failure to consider changes in factor supplies may result in wrong conclusions concerning the effect of an increased demand for skilled labor on the wage ratio if the supply of skilled labor has increased relative to the supply of unskilled labor.

In contrast to the Heckscher-Ohlin-Samuelson model, the dynamic intertemporal general equilibrium model in this paper examines a small open economy with three types of goods: exports, imports, and non-traded goods; three agents: firms, households and government; and two kinds of labor: skilled labor and unskilled labor. The production factor endowments, i.e. physical capital, skilled labor and unskilled labor, are allowed to vary over time in line with the optimising choices of the three agents. The model goes a step further than CGE models in that it shows not only the long-run transitions but also the short-run transitions of the endogenous variables. It shows that what happens in the long run may not be a good guide to what happens in the short run. A modelling framework is summarized as follows.[7]

A. Firms

Firms employ physical capital, skilled labor and unskilled labor to produce three types of goods. The firms sell these goods to households for consumption, to government for education capital investment, and to themselves for physical capital investment. There are three representative firms in the economy; they represent, respectively, the export sector, the import sector and the non-traded sector. Exports are a function of foreign income and the inverse of terms of trade. The capital accumulation in each sector depends on the rate of fixed capital formation and the rate of depreciation.

To initialize the model, it is assumed that sector 1 is relatively skilled-labor intensive, sector 2 is relatively unskilled-labor intensive and sector 3 is relatively capital intensive. It is also assumed that exports consist of good 1, imports consist of good 2 and that good 3 is non-traded. It follows from these assumptions that exports are relatively skilled-labor intensive and imports are relatively unskilled-labor intensive. Table 1 illustrates these characteristics.

Table 1. Sector Characteristic

Sector	Factor intensive	Trade
1	Skilled labor	Export
2	Unskilled labor	Import
3	Capital	Non-traded

[7] Refer to Chang (2000) for a detailed discussion of this model.

B. Households

Households supply unskilled labor to firms and skilled labor to both firms and government in return for wages. They also own the physical capital and earn financial dividends. Their income is used to finance goods consumption from firms and purchases of education from government. The opportunity cost of leisure is the forgone opportunity of working. To maximize their utility households distribute consumption optimally on both goods and leisure under their budget constraints.

The optimal net skill formation chosen by households is the fixed skill formation minus skill depreciation. Households' education spending depends on fixed skill formation and an adjustment cost function. The adjustment cost reflects the forgone production and relies on the ratio of fixed skill formation to skilled labor. If skilled labor is increasing, the adjustment cost is decreasing. This is plausible due to the spill over effect within the labor force. The elasticity of the skilled labor supply, with respect to the wage of skilled labor, is greater than zero since the supply of skilled labor is not fixed. It is less than infinity in the short run, because the transformation from unskilled to skilled labor is not free. Some skills are specific or patented, and training facilities are not always available. Hence, the supply of skilled labor is also not perfectly elastic in the long run. Technically, because of the endogenous wages and the leisure variable in households' utility function, the labor supply of both types has an endogenous ceiling in this framework.

C. Government

The government buys goods from firms and transforms them into education capital. The government hires skilled labor and uses this in conjunction with the education capital to produce education. By its collection of labor income tax and by selling education to households the government exactly finances its spending on education capital investment and skilled labor. That is, the budget is balanced. The role of government as an education supplier is essential. This model captures the reality of government supplying education in consideration of the beneficial externalities resulting from education. Total government investment on education capital is assumed to be exogenously controlled by the government. The accumulation of education capital is given by the total investment by government minus the depreciation.

To ensure the model is consistent in achieving general equilibrium, the rule of demand equal to supply is applied to both the goods and the factor markets. The full model in the steady state is shown in Appendix.

D. The Wage Differential in the Steady State

Due to the complexity of the above model, there is no reduced form that can be derived to present the wage differential. An expression, possibly the simplest, of the wage differential in the steady state is as follows,

$$W_s = W_u + [\, P_E / (1 - \tau)]\, [\theta + \theta \;\Phi\; \delta_s + \delta_s + (1/2)\, \Phi\, \delta_s^{2}\,] \qquad (1)$$

where W_s is the skilled wage, W_u is the unskilled wage, P_E is the price of education, τ is the tax rate, θ is the rate of time preference, Φ is the skill adjustment cost parameter and δ_s is the rate at which skill depreciates.

The expression of equation (1) is independent of the functional form of both the utility and the production function.[8] The equation provides a rigorous theoretical result for the wage differential. The relationship between the skilled and the unskilled wage depends on parameters, namely, the rate of time preference (θ), the depreciation rate of skill (δ_s) and the skill adjustment cost parameter (Φ); and on endogenous variables, namely, the tax rate (τ) and the price of education (P_E). A higher skill adjustment cost and a higher skill depreciation rate tend to boost the cost of skill formation, therefore leading to a higher skilled wage. The rate of time preference counts because an investment in skill formation takes time to repay. A larger time preference involves a larger adjustment cost for skill formation, therefore a patient household will expect a higher skilled wage. The education price and the tax rate are endogenous in this model. Theoretically, each endogenous variable in (1) can be solved and substituted by the exogenous variables and parameters, so implicitly the wage differential is a function:

[8] A detailed proof is available from the author.

$$\frac{W_s}{W_u} = f\,(A_{Q_i}, I_E^G, Y^* \mid parameters) \qquad (2)$$

where A_{Qi} is the technology variable for each sector, I_E^G is the education investment controlled by the government and Y^* is the foreign income which directly affects domestic exports.

The wage differential equation (1) illustrates both the importance and the transmission channel of education in the determination of the wage differential. This substantiates the inclusion of education in the debate on the wage differential, in addition to the traditional arguments of trade and productivity growth. The government, as an education supplier and tax collector, has the ability to control the wage differential to a certain extent. What matters in a general equilibrium outcome is the interactive effect of the education price and the tax rate. Simulation becomes necessary to explore the short- and long-run transitions of each endogenous variable and so establish the policy implications.

E. Simulation Results

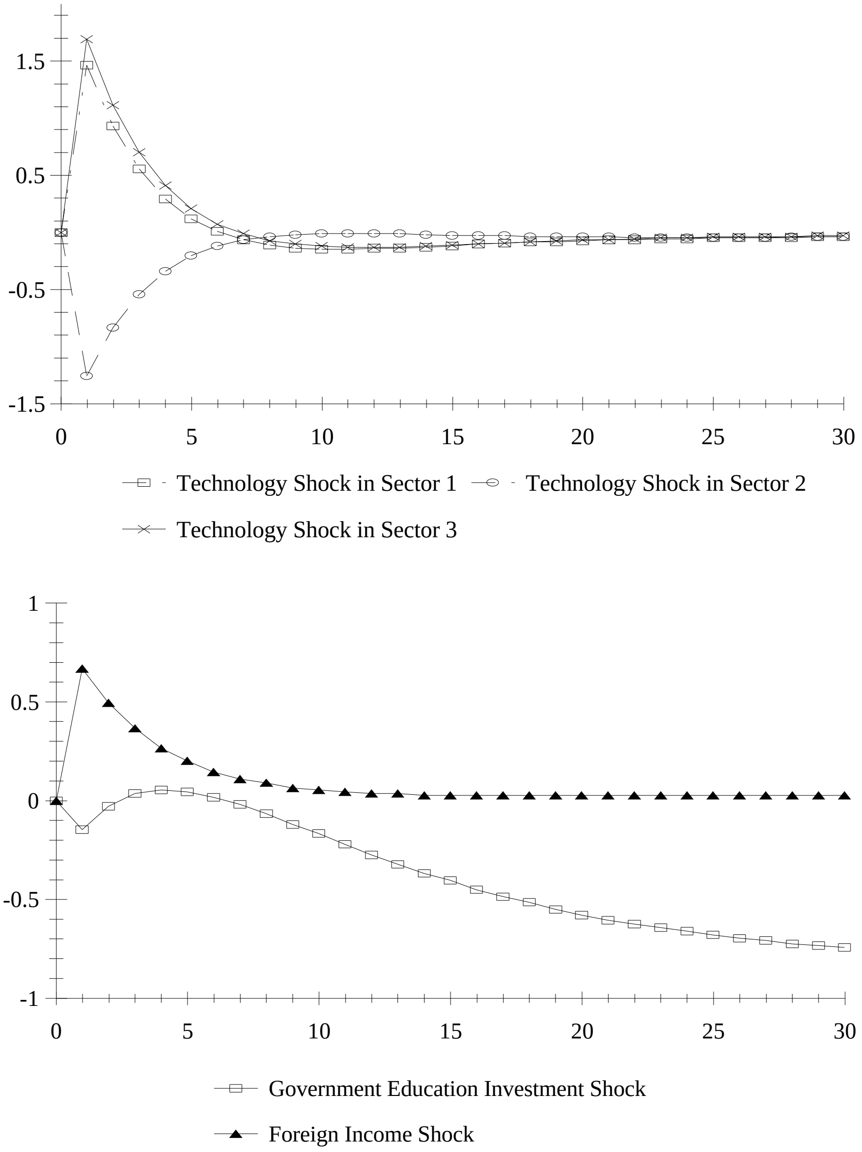
The main results from this model are that, in the long run, productivity growth and an increase in government education investment lessen the wage differential.[9] Generally speaking, increased education investment also lessens the wage differential in the short run, albeit with a fluctuation in the early stage. (The fluctuation occurs because the adjustment process of skill formation takes time and households make optimal choices between working and leisure.) Productivity growth, at most, raises the wage differential only in the short run: it may reduce the wage differential in the short run if productivity growth is biased towards unskilled labor. An increase in international trade increases the wage differential to a larger extent in the short run than in the long run.[10] These simulation paths are presented in Figure 3.

Intuitively, an increased demand for skilled labor resulting from a growth

[9] Productivity growth lessens the wage differential to a small but non-zero extent.

[10] In the long run, the wage differential is enlarged to a small but non-zero extent.

Figure 3. Simulation Results: Wage Ratio



in productivity or exports can eventually be filled in the long run as skill supply plays an important role in the wage determination.[11] In the short run, the created demand cannot be filled immediately due to the time required for skill formation. By using equation (1), the transitions are as follows. Productivity growth pushes down goods prices. This reduces the costs of government purchases and motivates the government to cut the tax rate, which, in turn, decreases the wage differential. If the government increases education investment, thereby decreasing the education price, it can cause diminution of the wage differential. While the international trade factor is not explicitly shown in equation (1), its effect is transmitted from production to wages through the education price. An increase in skill-intensive exports boosts the price of exports and the demand for skilled labor. This increases the demand for education and, hence, the price of education. Therefore, the wage differential rises.

From a theoretical perspective it is unconvincing that productivity growth raises the wage differential in the long run since skill formation eventually catches up with the progress of technology as long as the adjustment cost is affordable for the unskilled labor.

F. Sensitivity Test

Since there are three sectors with a different intensity of each factor, and five different shocks -technological progress in sectors 1, 2 and 3, government education investment and foreign income-, there are a total of thirty cases within this framework. Sector 1 is designated the export sector, sector 2 the import sector, and sector 3 the non-traded sector.

The key variable investigated is the change of the wage ratio in the steady state. The results, set out in Table 2, show that this model is fairly robust. In Table 2, the first column gives the combination of three sectors with different

[11] In the simulation, an aggregation of three sectors leads to a case of factor-biased productivity growth, a reason emphasized by Krugman and Laurence (1994) for enlarged the wage differential. However, the present model allows the skill demand and supply to determine the skilled wage whereas in their paper it is asserted that increased demand results in an increased wage.

intensities of inputs. Numbers stand for sectors and letters stands for inputs, for example 1U2S3K represents sector 1 as unskilled-labor intensive, sector 2 as skilled-labor intensive and sector 3 as capital intensive. The second to sixth columns respectively stand for an improvement in technology in sectors 1, 2 and 3, an increase in government education investment and an increase in foreign income. A minus sign (-) means a decreased wage differential and a plus sign (+) means an increased wage differential.

Table 2. Sensitivity Test: The Effect of Shocks on the Wage Differential

	Tech 1	Tech 2	Tech 3	Government *EDUN*	Foreign income
1S2U3K	-	-	-	-	+
1S2K3U	-	-	-	-	+
1U2S3K	-*	-*	-	-	-
1U2K3S	-	-	-	-	-
1K2S3U	-	-	-	-	-
1K2U3S	-	-	-	-	+

Note: * When the share of education capital in education production is equal to or greater than 0.5, the sign becomes +.

To summarize, on the one hand, the effect on the wage differential of an improvement in technology in any of the sectors or of increased government investment in education is insensitive to different combinations of sectors, that is, in each case the wage differential decreases in the long run, whereas, on the other hand, the effect on the wage differential of a foreign income shock that raises exports is sensitive to different combinations of sectors.

IV. Empirical Testing

This section demonstrates empirical tests for the theoretical results based on equation (2). The data set tested in this section is from several data sources. The monthly average wage data by education attainment comes from the

Manpower Utilization Survey, which is published by the Directorate-General of Budget, Accounting, and Statistics (DGBAS) of the Republic of China. As mentioned in Section II, the raw wage data have been manipulated to be a weighted average level. Government investment in education is proxied by the share of government expenditure on education, science and culture in GDP (*EDUN*). This measure, which reflects a broad definition of education investment, is from the CEIC Database, which is maintained by EconData Pty. Ltd.. The use of *EDUN* as a proxy is justified for the following reasons. During the past two decades, first, the Taiwanese government did not overwhelmingly target any specific education level: education expenditure per student at all levels increased to a similar extent.[12] Second, university and college levels took a rapidly growing share of total spending on education owing to more high school graduates entering the university level.[13] These two reasons make the following proposition plausible, for a case study of Taiwan, that more government education investment forms more skilled labor and thereby decreases the wage differential.

The proxies for productivity growth and international trade are, respectively, the annual growth rate of total factor productivity (*TFP*) and the share of net exports in GDP (*NETX*). These proxies are drawn from various issues of the Taiwan Statistical Data Book, published by the Council for Economic Planning and Development of the Republic of China.[14] Since the wage data are drawn from the whole economy, the *TFP* calculated from the combined industry (manufacturing, construction, and electricity, gas and water), agriculture and service sectors is an appropriate explanatory variable to use in testing the effect of productivity growth on the wage differential. Net exports, which are driven by foreign income (the shock tested in the theoretical model), measure an approximate net effect of international trade on the wage differential.

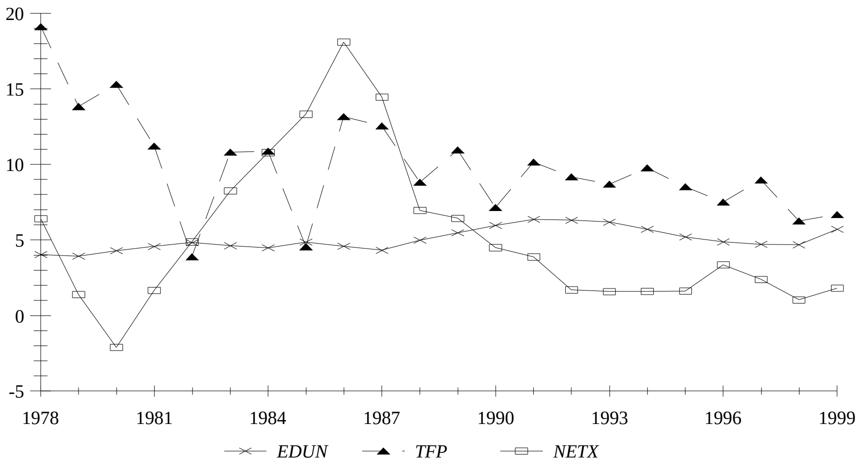
[12] Table 14-10 in Taiwan Statistical Data Book 2000.

[13] The rapidly growing share results from the policy of removing the government-imposed limit on the number of tertiary education institutes in Taiwan. Discussion of this issue is beyond the scope of this paper.

[14] Total factor productivity is the weighted average, by using the shares in GDP as the weights, of the annual growth rates of the industry, service and agriculture sectors.

The time series covers the period from 1978 to 1999. Figure 4 presents a graphical description of the variables *TFP*, *EDUN* and *NETX*. Due to it being a small sample, a Bootstrapping estimation is constructed for the robustness test. The wage differential is measured by W_s/W_u, that is, by the ratio of the average monthly wage with a college or above degree (skilled labor) to that with a degree from junior college or below (unskilled labor).[15] Based on the theoretical framework in the previous section, a long-run model and an Error Correction Model are established to demonstrate both the long run and the short-run effects of *TFP*, *EDUN* and *NETX* on the wage differential.

Figure 4. *EDUN*, *TFP* and *NETX* (Unit:%)



A. Unit Root Tests

Table 3 illustrates the results of Dickey-Fuller unit root tests.[16] Although the Dickey-Fuller test is known to have low power in testing for unit roots, especially when dealing with a small sample, it still provides suggestive results for the stationarity of time series. The Dickey-Fuller test for unit roots shows that *TFP* is I(0) and the other three variables, W_s/W_u, *EDUN*, *NETX*, are I(1).

[15] To focus on the conventional definition of college or above as skilled labor is plausible due to the systematic shift downwards of the wage ratio if junior college is included.

[16] Phillips-Perron unit root tests end with the same results.

Table 3. Unit Root Tests

	W_s / W_u	*TFP*	*EDUN*	*NETX*	C.V. 5%[1]
constant, no trend [2]					
A(1) = 0 t-test	-1.5	-3.4	-1.5	-1.3	-2.9
A(0) = A(1) = 0	1.2	6.0	1.5	0.9	4.6
constant, trend					
A(1) = 0 t-test	-1.9	-4.7	-1.6	-1.4	-3.4
A(0) = A(1) = A(2) = 0	1.3	7.6	1.1	0.7	4.7
A(1) = A(2) = 0	1.8	11.2	1.3	1.0	6.3
Conclusion	I(1)	I(0)	I(1)	I(1)	

Notes: [1] C.V. 5% means critical value at 5% significance level. [2] The Augmented Dickey-Fuller (ADF) regression equations in Shazam use the first-difference regressant with and without a time trend, where A(0) is the "drift" coefficient, A(1) is the coefficient of the tested variable with one lag, and A(2) is the time trend coefficient. The null hypothesis for the existence of unit roots is A(1) = 0.

B. Long-Run Model

Since the sample size is small, the power of the Dickey-Fuller unit root test is low and both I(0) and I(1) may be included as explanatory variables for this case. Hence, whether or not *TFP* should be included in the model is examined. The result shows that *TFP* is an insignificant long-run factor in the determination of the wage differential.[17] Therefore, the following long-run model is proposed and estimated to examine the long-run relationship.[18] Table 4 shows the estimation results.

[17] The theoretical model suggests that in the long run, *TFP* and international trade respectively have a small negative and a small positive effect on the wage differential. The empirical data can further determine their significance (or lack thereof) in the empirical model.

[18] This linear specification performs a better statistical significance in terms of a range of diagnostic testing than the non-linear specification with which the square terms of *EDUN* or/and *NETX* are embedded.

$$W_s / W_u = \beta_0 + \beta_1 EDUN + \beta_2 NETX + \varepsilon \qquad \varepsilon \sim \text{i.i.d. } N(0, \sigma_\varepsilon^2) \qquad (3)$$

Table 4. The Long-Run Model

Variable	Estimated coefficient	Standard error	t-ratio 16 d.f.	p-value	Elasticity at means
EDUN	-3.9574	1.7694	-2.2365	0.0375	-0.1178
NETX	1.0415	0.2536	4.1067	0.0006	0.0319
constant	184.1	9.4496	19.472	0.0000	1.0863
Durbin-Watson = 1.7985					
R-square adjusted = 0.8027					
Log of the likelihood function = -58.3249					

In the long run, trade and education have a significant effect on the wage differential. If government education investment increases by 1 per cent of GDP, the wage ratio drops by about 0.04 (2.34 per cent of the average wage differential over the period). If net exports increase by 1 per cent of GDP, the wage ratio rises by around 0.01 (0.61 per cent of the average wage differential over the period). This shows that, in the long run, government education investment has a larger effect on the wage differential than do net exports. Following from the theoretical simulation, this positive effect of net exports on the wage differential implies that Taiwan's exports are relatively skilled-labor intensive compared with imports. This result for the effect of net exports on the wage differential is consistent with the findings of Chan et al (1998) and is stronger than the findings of Chen and Hsu (2001).[19]

C. An Error Correction Model

The following ECM provides a case of *TFP* only affecting the wage differential in the short run. *EDUN* and *NETX* are included in both the short

[19] Using a full model regarding the Taiwanese economy as a whole, they find that net exports have a positive but insignificant effect on the wage differential.

and the long run. In equation (4), if the terms in the brackets are used this may not satisfy the regularity condition in the sense that these terms are I(1) while the left-hand side is I(0). Also, to avoid losing degrees of freedom, the bracketed terms are replaced by the residual from the long-run model. Table 5 illustrates the results of the estimation of equation (4) after correcting for both heteroskedasticity and autocorrelation using the Shazam program.

$$\Delta(W_s / W_u) = \beta_0 + \beta_1\ TFP_t + \beta_2\ \Delta EDUN_t + \beta_3\ \Delta NETX_t + \qquad (4)$$

$$+ \gamma [(W_s / W_u)_{t-1} - \delta_2\ EDUN_{t-1} - \delta_3\ NETX_{t-1}] +$$

$$+ \omega_t \qquad \omega_t \sim i.i.d.\ N(0, \sigma^2_\omega)$$

Table 5. An ECM without Long-Run Effect of TFP and Bootstrapping Estimation

Variable	Estimated coefficient	Standard error	t-ratio 16 d.f.	p-value	Bootstrapping means
TFP	-0.2472	0.0620	-3.9877	0.0011	-0.2471
Δ*EDUN*	-3.9247	0.6333	-6.1973	0.0000	-3.9298
Δ*NETX*	1.1644	0.0715	16.279	0.0000	1.1671
RESIDUAL	1.2249	0.0404	30.295	0.0000	1.2253
constant	2.3200	0.8151	2.8462	0.0117	2.2856

Durbin-Watson = 2.0293

R-square adjusted = 0.9653

Log of the likelihood function = -28.2702

This ECM estimation results in a fairly good match to the simulation results in Section III. In the short run, if the total factor productivity growth increases by 1 percentage point (e.g. from 6 per cent to 7 per cent), the wage ratio drops by about 0.0025 (0.15 per cent of the average wage differential over the period). Corresponding to the theoretical simulation, the effect of total factor productivity in decreasing the wage differential implies that Taiwan’s productivity growth in terms of the whole economy is unskilled-biased. If government education investment increases by 1 per cent of GDP, the wage ratio drops by about 0.039 (2.32 per cent of the average wage differential over the period). If net exports increase by 1 per cent of GDP, the wage ratio rises by around 0.012 (0.69 per cent of the average wage differential over the period). Taiwan's exports have shifted from having a high degree of labor intensity to having a medium or high degree of technology intensity. In line with the upgrade of technology, greater skilled labor intensity is also embedded in exports. Its imports have shifted from having a low degree of labor intensity to having a high degree of labor intensity. Incorporating these two facts, the empirical result of international trade raising the wage differential is consistent with the large country cases. Comparing these results with those in the long run, government education investment has a relatively smaller effect in decreasing, and net exports have a relatively larger effect in increasing, the wage differential in the short run. These results are consistent with the theoretical results, which substantiate that skill formation takes time, and they add a new dimension to the results of the existing research.

Since these empirical data involve a small sample size, a Bootstrapping procedure (Efron, 1979) with a 2000 random re-sampling replication is used to test the robustness of the estimation. The Bootstrapping estimation is shown in the last column of Table 5. The mean of each variable is fairly close to its estimated coefficient in the above ECM model. This supports the validity of the estimation.

V. Conclusion

This paper portrays the profile of Taiwan's wage differential and employs an error correction model, which can perform tests in both the short and long run, to examine the effects of international trade, productivity growth and education investment on Taiwan's wage differential. Whether or not junior college graduates are designated to skilled or unskilled labor, there is no convincing evidence of a growing wage differential in Taiwan over the past two decades.

That education could be an important determinant of the wage differential is substantiated by both the theoretical model and the empirical data. Education investment takes time to have its full effect, therefore the empirical data show smaller decreases in the wage differential in the short run than in the long run. In the long run, if government education investment increases by 1 per cent of GDP, the wage ratio drops by about 2.34 per cent due to more skilled labor being available in the economy. International trade is also a significant determinant of the wage differential. If net exports increase by 1 per cent of GDP, the wage ratio rises by around 0.69 per cent in the short run and by around 0.61 per cent in the long run. Productivity growth has a significant influence on the wage differential in the short run, but may have only a minor effect in the long run. If total factor productivity growth increases by 1 percentage point, the wage ratio drops by about 0.15 per cent in the short run. This study thus points out that the short-run effects are different from the long-run effects, adding a new dimension to the existing research.

An inference that the skill adjustment cost in Taiwan is low enough to allow unskilled labor to be transformed into skilled labor (when skilled labor is required) can be made for the Taiwanese economy. By pointing out the importance of the skill adjustment cost in the determination of the wage differential, this paper proposes a new angle for future researches. Different countries face different affordable skill adjustment costs. Even within a country, the skill adjustment cost may vary over time as a result of other changes in the economy.

Appendix

Table A.1. The Theoretical Model in the Steady State

	Equations	
Q_i	$=$	$A_{Qi}\, K_i^{\alpha i}\, L_{s,i}^{F\,\beta i}\, L_{u,i}^{1-\alpha i-\beta i}$
$J_{t,\,i}$	$=$	$\delta_i K_{t,\,i}$
I_i	$=$	$J_i\,(1 + \phi_i\,\delta_i/2)$
$Q_{i,\,Ls}$	$=$	W_s/P_i

Table A.1. (Continued) The Theoretical Model in the Steady State

Equations		
$Q_{i,Lu}$	=	W_u/P_i
λ_i	=	$1+\phi_i\delta_i$
Q_{Ki}	=	$(r+\delta_i)\lambda_i - \phi_i\delta_i^2/2$
$P_{2,t}M_t$	=	$P_{1,t}X_t$
X	=	$(P_2/P_1)^{\rho}\,Y^*$
0	=	$r_t F_t + (1-\tau_t)\,[(W_s/P_2)\,L_{s,t} + (W_u/P_2)\,L_{u,t}] - [(P_1/P_2)\,C_{1,t} + C_{2,t} +$
		$+ (P_3/P_2)\,C_{3,t} + (P_{E,t}/P_2)\,S_{E,t}]$
$J_{s,t}$	=	$\delta_s L_{s,t}$
F_t	=	$(P_1/P_2)\,\lambda_{1,t}K_{1,t} + \lambda_{2,t}K_{2,t} + (P_3/P_2)\,\lambda_{3,t}K_{3,t}$
$S_{E,t}$	=	$J_{s,t}\,(1+\Phi\,\delta_s/2)$
l_t	=	$T - L_{s,t} - L_{u,t}$
U_{Ci}	=	$(P_i/P_2)\,\mu_1$
$U_{Lu,t}$	=	$-\mu_1\;(1-\tau)\;W_u/P_2$
μ_2	=	$\mu_1\,P_E\;(1+\Phi\,\delta_s)/P_2$
r_t	=	θ
U_{Ls}	=	$(\theta+\delta_s)\,\mu_2 - \mu_1\,[(1-\tau)\,W_s + P_E(\Phi\,\delta_s^2)/2]/P_2$

Table A.1. (Continued) The Theoretical Model in the Steady State

	Equations	
S_E	=	$K_E^{\xi}\ L_s^{G^{1-\xi}}$
I_E^G	=	$\delta_E\ K_E$
$I_{E,t}^G$	=	$(P_1\ I_{E,1}^G + P_2\ I_{E,2}^G + P_3\ I_{E,3}^G) / P_E^G$
$P_E^G\ I_E^G + W_s\ L_s^G$	=	$\tau\ (W_s\ L_s + W_u\ L_u) + P_E\ S_E$
$Q_{1,t} - X_t$	=	$C_1 + I_{E,1}^G + I_1$
$Q_{2,t} + M_t$	=	$C_2 + I_{E,2}^G + I_2$
$Q_{3,t}$	=	$C_3 + I_{E,3}^G + I_3$

Note: Notation: (subscript i = 1, 2, 3 stands for Sector 1, 2, and 3). Q: Production; A: Technology; K: Capital; L_s^F: Skilled labor hired by firms; L_s^G: Skilled labor hired by government; L_s: Total skilled labor; L_u: Unskilled labor; J: Fixed capital formation; I: Capital investment; W_s: Skilled wage; W_u: Unskilled wage; P: Price; M: Imports; X: Exports; Y^*: Foreign Income; F: Financial asset; C: Consumption; S_E: Amount of education buying; J_s: Fixed skill formation; I_E: Household's education investment; l: Leisure; T: Time constraint; U_Z: Marginal utility of Z; P_E: Price of education; r: Interest rate; K_E: Education capital; I_E^G: Government education investment; P_E^G: Weighted price index; τ: Tax rate; α, β: Input shares in goods production function; δ: Depreciation rate of capital; ϕ: Adjustment cost parameter of capital investment; λ: Shadow price of capital; ρ: Parameter; δ_s: Depreciation rate of skill; Φ: Skill adjustment cost parameter; μ_1: shadow price of financial asset; μ_2: shadow price of skill; θ: Rate of time preference; ξ: Input share in education production function; ε: Weight of pooled price index; δ_E: Depreciation rate of education capital.

References

Bhagwati, J. and V.H. Dehejia (1994), "Freer Trade and Wages of the Unskilled- Is Marx Striking Again?," in J. Bhagwati and M.H. Kosters, eds., *Trade and Wages: Leveling Wages Down?*, Washington, D.C., AEI Press.

Bound, J. and Johnson, G. (1992), "Changes in the Structure of Wages in the 1980s: An Evaluation of Alternative Explanations," *American Economic Review* **82**: 371-392.

Chan, V.L., Chen, L.T. and S.C. Hu (1998), "Implications of Technology and Education for Wage Dispersion: Evidence from Taiwan," in G. Ranis, S.C. Hu and Y.P. Chu, eds., *The Political Economy of Taiwan's Development into the 21st Century*, Cheltenham, Edward Elgar.

Chang, H.C. (2000), "Wage Differential, Trade, Productivity Growth and Education," Working Papers in Trade and Development 2000/1, The Australian National University.

Chen, B.L. and M. Hsu (2001), "Time-Series Wage Differential in Taiwan: The Role of International Trade," *Review of Development Economics* **5**: 336-354.

Council for Economic Planning and Development, Various Years, *Taiwan Statistical Data Book*, Taipei, Republic of China.

Deardorff, A.V. and D.S. Hakura (1994), "Trade and Wages -What Are the Questions?," in J. Bhagwati and M.H. Kosters, eds., *Trade and Wages: Leveling Wages Down?*, Washington, D.C., AEI Press.

Directorate-General of Budget, Accounting, and Statistics (DGBAS), Various Years, *Manpower Utilization Survey*, Taipei, Republic of China.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics* **7**: 1-26.

Katz, L.F. and D.H. Autor (1999), "Changes in the Wage Structure and Earnings Inequality," in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics 3*: 1463-1555, New York and Oxford, Elsevier Science, North-Holland.

Katz, L.F. and K.M. Murphy (1992), "Changes in Relative Wages, 1963-1987: Supply and Demand Factors," *Quarterly Journal of Economics* **107**: 35-78.

Lawrence, R.Z. and M.J. Slaughter (1993), "Trade and U.S. Wages: Giant Sucking Sound or Small Hiccup?," *Brookings Papers on Economic Activity*, Microeconomics: 161-210.

Leamer, E.E. (1994), "Testing Trade Theory," in D. Greenaway and L.A. Winters, eds., *Surveys in International Trade*, Oxford, Blackwell Publishers.

Tsay, C.L. (1995), "Labor Migration in Asia: Taiwan," *ASEAN Economic Bulletin* **12**: 175-190.

Wood, A. (1994), *North-South Trade, Employment and Inequality, Changing Fortunes in a Skill-Driven World*, Oxford, Clarendon Press.

Journal of Applied Economics, Vol. VI, No. 1 (May 2003), 49-72

ON THE VALUATION OF COMPANIES WITH GROWTH OPPORTUNITIES

JOSÉ PABLO DAPENA[*]
Universidad del CEMA

Submitted March 2001; accepted December 2001

Each company faces day to day investment opportunities. Just by staying in business the company is taking a decision of reinvesting capital. These opportunities have to be fairly valued to overcome misallocation of resources. A project with high growth opportunities requires high reinvestments to take full advantage of them until it reaches its mature stage. These investments can be seen as a succession of call options on future growth. When a company with such prospects is valued using the discounted cash flow technique and growth is taken implicitly in the growing cash flows and the residual value, the value thus obtained will be higher than the true one (under certain circumstances). Technology advances and the effects of globalization create enormous growth opportunities, and so misvaluation risks are higher.

JEL classification codes: G 12
Key words: real options, valuation, contingent claims valuation

I. Introduction

For decades there has been a fruitful use of the method of Discounted Cash Flow and Net Present Value (henceforth DCF and NPV respectively) to value and evaluate business projects and investment opportunities.[1] They have become standard tools that any financial analyst and manager should manage

[*] I gratefully acknowledge helpful comments from Dr. Rodolfo Apreda and Dr. Edgardo Zablotsky. I also want to thank participants at the XXV Annual Meeting of the Asociación Argentina de Economía Política, where this paper was presented. Correspondence should be addressed to: jd@cema.edu.ar.

[1] For a more detailed analysis see the initial chapters of "Corporate Finance," by Stephen Ross.

and master to value investment prospects. The DCF works by discounting the expected stream of cash flows using a risk adjusted rate of return.[2] Even though this form of DCF became of great utility, it could not be used to value assets whose payoff are asymmetrical, like options and other derivatives. The breakthrough to the correct valuation of such contracts was made by the contributions of Black and Sholes (1973) and Merton (1973) with the derivation of the valuation formula under certain assumptions, and followed by Cox, Ross and Rubinstein (1979) and the development of the risk neutral approach to valuation.[3]

Since these developments practitioners in finance found themselves equipped with two powerful tools to value streams of cash flows, the standard discounted cash flow technique and the option pricing methodology.[4] Myers (1977) was the first to note that the value of a firm is composed of a stream of cash flows for whom both tools can be used to reflect its value. He showed that the value of any firm is composed of two building blocks, the value of assets in place and the value of growth opportunities. Dixit and Pindyck (1994) showed in a comprehensive book how uncertainty can modify investment rules taken for granted, and how the rule of “invest in projects with positive NPV” does not strictly obtain (in a sense that projects with negative NPV are nevertheless undertaken) for some cases. More recently, the work of Trigeorgis (1988, 1997) and Kulatilaka (1992, 1995) showed how traditional DCF analysis fails to take into account the value of options embedded in projects, prompting undervaluation, and providing rationality to the fact that projects with negative NPV are nevertheless undertaken by companies.[5] The basic idea developed is that the use of traditional DCF to obtain NPV does not consider the flexibility inherent in some projects the management has to react

[2] See next section for the analysis.

[3] See also Mason and Merton (1985).

[4] Although the option pricing technique is a particular form of discounting cash flows, we shall use the term traditional DCF to refer to discounting the expected stream of cash flows using a risk adjusted rate of return.

[5] The negative net present value is outwheigthed by the positive value of the options embedded.

to either favorable or unfavorable conditions, and hence does not include the value of such flexibility. This evidence has prompted a lot of academic work to show undervaluation of projects due to neglected embedded options.[6]

It is the purpose of this paper to show the other side of the coin, those situations where the growth on the cash flows of the project are subject to reinvesting, which in turn is contingent on favorable events. In this case, the cash flow is valued using the traditional DCF technique, and as it is the objective of this paper to show that, under certain conditions, the valuation thus obtained tend to overvalue the true value of the stream of cash flows.

II. Valuation Techniques[7]

On this section we shall state the basic assumptions governing our world, and a brief revision of the conditions underlying the different valuation methods.[8]

A. Assumptions[9]

The following assumptions will be made to make the world more tractable: (a) the typical investor is risk averse, which means she requires a premium to hold assets with uncertain payoffs, (b) capital markets are complete, which means there is a price to be paid to obtain insurance against any state of the nature, (c) the information set is the same for all investors, meaning information is symmetric, (d) growth options embedded in projects take the form of European derivatives, where early exercise is not allowed, where this assumption will help structure the problem in a simple way, (e) the risk free rate is non-stochastic and given, which is a derivation of the assumption of

[6] See for example Kulatilaka (1992).

[7] We do not consider other methods like relative valuation (comparables) though we acknowledge their existence.

[8] The description of the two methods is adapted from the work mentioned in the introduction.

[9] These assumptions are not far more restrictive than those of the Capital Asset Pricing Model or the Contingent Claim Analysis.

complete capital markets, (f) the value of the company is unaffected by the capital structure, so there is no opportunity of creating value by changing the capital structure (in other words, the Modigliani-Miller theorem holds true), (g) the value of the business in each state of the nature is known, which in turn means there is no risk in assessing the payoffs in each state of the nature, (h) there is an appropriate way of obtaining the risk-adjusted rate of return properly reflecting risk preferences of investors,[10] (i) the probabilities of each state of the nature are known, and (j) in a binomial world when moving the value of probabilities, volatility changes. We shall ignore this effect on the risk-adjusted rate of return.

B. The Traditional Methodology

The traditional method accounts for the calculation of the expected value of future cash flows, discounting it using a risk adjusted rate of return,[11] intended to show the preferences towards risk of the average investor. In terms of a discrete distribution of probabilities, the present value of a one period project can be shown to be

$$V_t = \Sigma \frac{\Sigma \ p_{i,\ t+1} \ V_{i,\ t+1}}{(1+k)^{t+1}} \tag{1}$$

where $V_{i,\ t+1}$ represents the values the project or the firm can undertake in each state of the nature i at date t + 1 (from the cash flows it generates), p_i accounts for the likelihood of each state of the nature, k is the equilibrium risk adjusted rate of return from t to t +1.

[10] For example, the assumptions of the Capital Asset Pricing Model hold true.

[11] To the purpose of obtaining the appropriate rate for equity, a standard Capital Asset Pricing Model (CAPM) of the form, $k = E(R_i) = r_f + b_i(E(R_m) - r_f))$, can be used, where the left hand side represents the expected return the project has to earn, and the right hand side accounts for two terms, r_f for the risk free rate, and a risk premium. According to the model, in equilibrium the investor pays only for the risk he cannot diversify by himself. It is assumed that value is independent of the capital structure, so there is no point on differentiating between equity and debt.

The value thus obtained is the value of the stream of cash flows, which is then compared with the required initial outlay in order to decide whether the opportunity is worth to be undertaken. If the difference between both (value minus cost of investment) is positive, the project is pursued.[12]

C. Contingent Claim Analysis

Alternatively, in a complete capital market an investor can pay a price π_i at time t to obtain a pure asset, which pays a dollar at t + 1 should state i of the nature happen and zero otherwise.[13] Investors wanting to ensure one dollar in every state of the nature will have to buy a complete set of pure assets paying for it the sum of the prices of each pure asset ($\Sigma\, \pi_i$). The portfolio thus obtained will have the property of being riskless (the payoff of such a portfolio is the same regardless of the state of the nature), hence in equilibrium and to rule out arbitrage opportunities, the return of such a portfolio has to be equal to the risk free return. We label the risk free rate by r, thus $\Sigma\, \pi_i = 1/(1+r)$. Therefore, in equilibrium an asset that pays or has a value of V_i dollars in the state of the nature i and zero otherwise has to be worth $\pi_i\, V_i$. We have that the value V of such a project or firm is shown to be:[14]

$$V_t = \Sigma\, \tilde{p}_i \;\; V_{i,t+1} \;\; \frac{1}{(1+r)} \tag{2}$$

In other words, the value is the expected value of the payoffs using a synthetic probability distribution, discounted at the risk free rate. It can be easily seen that this new probability distribution satisfies all the requirements of any probability distribution: non-negative values, the sum of all at a certain time adding up to one, etc. We have valued the project using the risk free rate in the discount factor, just as if the investor was risk neutral. Nevertheless, it is shown that the value of the project V_t obtained is the same under the two alternatives.

[12] This is the NPV methodology.

[13] See for example the description by Varian (1992), chapter 20, pp. 448-452.

[14] See Appendix 1.

III. Growth Options

A. Flexibility on Decisions

Allocating resources in a company does not imply a rigid plan of activities, but a set of decisions conditional upon new information arriving, so decisions are sequential and cannot be fully planned in advance. This means decisions have to be taken as uncertainty unfolds, at the right moment. In these situations the manager needs not to take a decision until she counts with more information. As long as this flexibility does not cause a loss to the company, it has a positive value. These decisions the manager faces when allocating resources can be grouped into the following broad categories:[15] growth decisions, contraction (or even abandonment) decisions, and delay decisions. In all cases the company faces options that can be exercised only if events turn out to be favorable.[16] This reflects the right (not the obligation) the management has. This flexibility (or the options it implies) has value, and it is non-trivial for the value of the company.[17] In this paper I shall focus the analysis on reinvestment as a growth option, its structure and valuation.

B. Growth Decisions

A company can face a project which allows, in case events turn out to be good and circumstances are appropriate, to expand further. Even though this decision is not taken at the outset, the current value of the firm should reflect this option. Growth decisions that a manager can face are: expand business vertically (buy out or set up business within the value chain), expand business laterally (buy out or set up business not directly related with the core business), and expand the business (gain market share) by means of scope or scale.

[15] Adapted from Kulatilaka (1992).

[16] Otherwise the company can let the option expire and not exercise it.

[17] For example, two companies identical in everything but with a particular customer portfolio each, which allows one company to cross sell more products or services should market conditions turn favorable, cannot be worth the same.

Continuing with the valuation structure described above, we assume that in a particular state of the nature j at t + 1, the investor has the opportunity to undertake further investments with expected cash flows of n times the value of the project or firm at this moment (n $V_{i, t+1}$) by paying a cost K. This means the investor will pay the cost K only if n $V_{j, t+1} \geq$ K, or n $V_{j, t+1}$ - K ≥ 0.[18] If this inequality does not hold, the investor would be paying more than what the asset is worth. It can be seen that the investor would buy the asset (exercise her option to expand) only in those states of the nature where V_{t+1} is sufficiently high. In formula, the payoff or value of business in each state of the nature becomes

$$V_{i,\, t+1} + Max\ (n\ V_{i,\, t+1} - K,\ 0) \tag{3}$$

and the current value of business is thus (we shall label the current value of this asset $V_{t, A}$),

$$\mathrm{V}_{\mathrm{t,\ A}} = \frac{\Sigma\ p_i\ \ (V_{i,\ t+1} + Max\ \ (n\ V_{i,\ t+1} - K,\ 0))}{1+k} \tag{4}$$

This value (as shown before), can also be obtained using the contingent claim analysis or risk neutral valuation from (2). Now we shall label the value obtained by this method $V_{t, B}$

$$\mathrm{V}_{\mathrm{t,\ B}} = \frac{\Sigma\, \tilde{p}_i\ \ (V_{i,\ t+1} + Max\ \ (n\ V_{i,\ t+1} - K,\ 0))}{1+r} \tag{5}$$

where synthetic (or risk neutral) probabilities derived previously are used.

Throughout this paper we shall demonstrate that growing cash flows for business with growth opportunities require investing needs until they reach their mature stage, and this investment needs are growth options which must be correctly valued. The mature stage, used to value the business, implies exercising a succession of call options (the reinvesting) which must be valued according to their nature, and hence we will see that (4) overvalues the true

[18] We avoid the analysis of agency problems between managers and shareholders.

value of the business. Should this hypothesis be verified, it would mean that for some cases traditional valuation methodology has to be adjusted to reflect the overestimation.

Proposition: If the growing cash flows of a project are used to value it using DCF, and the growth on the cash flows involves reinvesting to attain them and to achieve a mature stage, the value of the project thus obtained will include the results of growth options already exercised through reinvesting, and the result will be an overvaluation of the true value of the project. The result is valid as long as the expected rate of return is greater than the risk free rate (the risk premium is positive).

Proof: (Two States of the Nature, One Period Model) Consider the simplest case, where we have two states of the nature at t+1, and the project value V can adopt two possible values, one for each state i. Assume there exists a risk free asset which pays a return of r. The likelihood of state 1 is given by p, while likelihood of state 2 is the complement 1 - p. According to the traditional method of valuation showed in (1), an asset of such features would be worth

$$\mathrm{V_t} = (p_1\ V_{1,\ t+1} + p_2\ V_{2,\ t+1})\ \frac{1}{1+k}$$

where k is a representative risk adjusted rate of return. Consistently with Appendix 2, we can find a synthetic probability $\tilde{p}$ based on the values V_1 and V_2, through which we obtain an expected value of V at t + 1. Discounting this expected value by using the risk free rate, the same value V_t derived by traditional methodology obtains.

This probability distribution based on $\tilde{p}$ comes out from setting the return of the asset equal to the risk free return, and changing the density mass of the probability distribution at each point of the possible values V at t + 1. The probabilities thus obtained are consistent with the current or spot value of the asset.

Armed with this synthetic probability, V_t is obtained by taking the expected value and discounting it to the risk free rate of return. As it was shown, the value V_t remains the same under the two methodologies, but in the second case the value is obtained as if the investor was neutral to risk.

We now capture the random structure of V from the parameters V_1,V_2 and (1 + r), which in turn are used to obtain the set of synthetic probabilities $\tilde{p}$ consistent with V_t.

Suppose that the future value involves growth through reinvesting, so there is a growth option embedded. As it was put as example before, the investor has the right to pay a cost of K to seize n times the value of V at t + 1 (we shall assume that in state 1 (nV) is greater than K, while in state 2 it is smaller), to make the manager exercise his option only in one state of the nature.[19] The asset's payoff then becomes

$$V_{i,t+1} + Max\ (n\ V_{i,t+1} - K, 0) \qquad \text{for i = 1, 2.} \qquad (6)$$

In state 1 we have, $V_{1,t+1} + (n\ V_{1,t+1} - K)$, while in state 2 the payoff is $V_{2,t+1}$. Given that the payoff in state 2 is the same under the two methods of valuation, for the sake of the comparison we can leave it aside and concentrate on the payoff in state 1. Under the traditional method of valuation, the value of the project including the expansion options would be

$$V_{t,\ A} = \frac{\Sigma\ p_i\ \ (V_{i,\ t+1} + Max\ \ (n\ V_{i,\ t+1} - K,\ 0))}{1+k} \qquad (7)$$

which for two states of the nature is

$$V_{t,\ A} = [p\ \ (V_{1,\ t+1} + n\ V_{i,\ t+1} - K) + (1-p)\ \ V_{2,\ t+1}]\ \frac{1}{1+k} \qquad (8)$$

rearranging terms we get

$$V_{t,\ A} = \frac{p}{1+k}\ (V_{1,\ t+1} + n\ V_{i,\ t+1} - K)\ +\ \frac{(1-p)}{1+k}\ V_{2,\ t+1} \qquad \text{and,}$$

[19] Otherwise would not be an option given it is exercised anyway.

$$V_{t,\ A} = \frac{p}{1+k}\ V_{1,\ t+1} \ + \ \frac{p}{1+k}\ (n\ V_{i,\ t+1} - K) \ + \ \frac{(1-p)}{1+k}\ V_{2,\ t+1}$$

making use of what we know about the value V$_t$, we notice that the structure of value is equal to the original value of the business plus the expansion option

$$V_{t,\ A} = \frac{p}{1+k}\ V_{1,\ t+1} \ + \ \frac{(1-p)}{1+k} V_{2,\ t+1} \ + \ \frac{p}{1+k}\ (n\ V_{i,\ t+1} - K)$$

being the first two terms equal to V$_t$

$$V_{t,\ A} = \ V_t + \frac{p}{1+k}\ (n\ V_{i,\ t+1} - K) \tag{9}$$

On the other hand, by using the risk neutral or contingent claim valuation method derived previously, we have

$$V_{t,\ B} = \ \Sigma\, \tilde{p}_i (V_{i,\ t} + 1 + Max\ \ (n\ V_{i,\ t+1} - K))\ \ \frac{1}{(1+r)} \tag{10}$$

which for the case of two states of the nature is given by

$$V_{t,\ B} = \ [\tilde{p}\ \ (V_{1,\ t+1} + n\ V_{1,\ t+1} - K) \ + \ (1-\tilde{p})\ V_{2,\ t+1}]\ \ \frac{1}{(1+r)} \tag{11}$$

following the same procedure of rearrangements of terms we have

$$V_{t,\ B} \ = \ \frac{\tilde{p}}{(1+r)}\ V_{1,\ t+1} \ + \ \frac{\tilde{p}}{(1+r)}\ (n\ V_{1,\ t+1} - K) \ + \ \frac{(1-\tilde{p})}{(1+r)}\ V_{2,\ t+1}$$

$$V_{t,\ B} \ = \ \frac{\tilde{p}}{(1+r)}\ (V_{1,\ t+1} + n\ V_{1,\ t+1} - K) \ + \ \frac{(1-\tilde{p})}{(1+r)}\ V_{2,\ t+1}$$

which according to our initial results can be written as

$$V_{t,\ B} \ = \ \frac{\tilde{p}}{(1+r)}\ V_{1,\ t+1} \ + \ \frac{(1-\tilde{p})}{(1+r)}\ V_{2,\ t+1} \ + \ \frac{\tilde{p}}{(1+r)}\ (n\ V_{1,\ t+1} - K)$$

We observe that again the value of the business is equal to the original value plus the growth or expansion option

$$V_{t,\ B} \ = \ V_t \ + \ \frac{\tilde{p}}{(1+r)} \ (n\ V_{1,\ t+1} - K) \tag{12}$$

comparing values for business obtained from each method ((9) and (12)), and simplifying for those terms equal in both derivations, we are left with the following simplified formula for traditional or DCF valuation

$$\frac{p}{(1+k)} \ (n\ V_{1,\ t+1} - K) \tag{13}$$

while the corresponding for risk neutral valuation is

$$\frac{\tilde{p}}{(1+r)} \ (n\ V_{1,\ t+1} - K) \tag{14}$$

given that the second factor of the multiplication is the same for both, we can drop it off for comparison purposes and concentrate on the first. If a univocal relationship is established between both, we are done. To this purpose, we make use of the components of any risk adjusted discount rate coefficient (1 + k). It is formed by the risk free factor (1 + r) times a risk premium (1 + θ)

$$(1 + k) = (1 + r)\,(1 + \theta) \tag{15}$$

Now we are allowed to make the last simplification. The risk free coefficient is present in both terms, so it can be dropped, then the comparison becomes

$$p\,/\,(1 + \theta) \ \ vs. \ \ \tilde{p} \ \ \text{or rearranging} \ \ \ p \ \ vs.\ \tilde{p} \ \ (1 + \theta) \tag{16}$$

if the first term in (16) is greater, it would mean that valuation of growth options by traditional DCF method overestimates the true value of the expansion opportunity. To prove this we use a basic axiom of the probabilistic theory, which says “...the probability is a non-negative number non-greater than 1."[20] Given that there is nothing in our derivation that can violate the axiom (the synthetic probability distribution comes out from a redistribution of mass at each point), and assuming the risk premium θ is positive[21] (being a parameter we can take it for given), $\tilde{p}$ can never be greater than p (if it was the case, and provided that we do not specify a specific value for this probability, we can always choose a value for $\tilde{p}$ to get a p greater than one, which in turn violates the axiom, so the relation must hold for every p and $\tilde{p}$. Hence, if the risk premium is positive for the underlying asset, the first term is always greater than the second, and the traditional method of valuation overestimates the true value of the growth option.

C. Extension of the Analysis from two States to n States of the Nature and to Continuous Time.

Having demonstrated the existence of overvaluation for the simple case of two states of the nature, we extend the framework to n states of the nature, where the random behavior of the variable is assumed to follow a binomial distribution with probability of success (upward movement) p, and n states of the nature. The maximum value that V can reach will have a probability of p^n associated, while the probability associated with the lowest value will be $(1 - p)^n$. For any value of V which requires j upward movements out of n possible, the probability associated will be $B\ (n; j; p) = C_j^n\ p^j\ (1 - p)^{n - j}$ where B denotes the binomial distribution.

Under the risk neutral valuation, the set of values V can adopt does not change, only does the density associated to each value, changing the mean of the distribution and adjusting it to the risk free return. As we saw in Section II, both methods give the same valuation for the underlying variable. The probability distribution thus obtained is of much help to value the options embedded in the project. We have to multiply each option payoff by its corresponding risk neutral probability, to obtain its expected, and then discount it to the risk free rate, obtaining the correct expected value. If we assume growth options are exercised when things go well, and we know that the true

[20] Mendenhall, Beaver and Beaver (1998), chapter 2, pp. 27-28.

[21] From our assumptions about risk preferences of the typical investor.

probabilities are greater for these states than their risk neutral counterpart, their complement for low value states will be smaller,[22] hence the inequality is reversed for low state values of the project. The demonstration is given by taking the upper bound, so that j = n, the true probability of this state or value would be

$$B\ (n; j; p) = C_n^{\,n}\ p^n\ (1 - p)^{n - n}\ =\ p^n \tag{17}$$

while the risk neutral would be

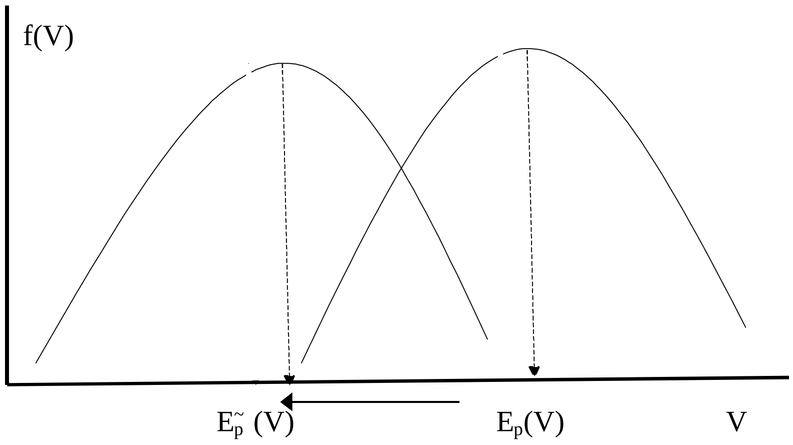
$$B\ (\ n; n; \tilde{p}) = C_n^{\,n}\ \tilde{p}^n\ (1 -\ p)^{n - n}\ = \tilde{p}^n \tag{18}$$

knowing that $p^{\sim}$ is smaller than p, any increasing monotonic transformation has to respect the inequality, so it can be said the following inequality holds, if $p \geq \tilde{p}$, then $p^n \geq \tilde{p}^{\,n}$. Both probability distributions have to integrate to one, so the excess in the upper side has to be offset by a diminution on the value of probabilities for low values of the underlying variable, so the inequality is reversed for such values $p \geq \tilde{p}$, then $(1 - p)^{\,n} \leq (1 - \tilde{p})^{\,n}$. When extending the framework to a continuous distribution, the binomial approximates the normal distribution as n → ∞, where the effect can be seen better on Figure 1, where V is the value of the company, f(V) is the density function (assumed normal), and it is seen that there is a redistribution of mass to change the expected value, which is less for the risk neutral distribution under a positive risk premium.

High values tend to have lower probabilities now. It can be seen clearly the effect of changing from the true distribution to a synthetic distribution when the risk premium is positive. It can be observed there is a redistribution of mass in the probability distribution to change and reduce the first moment of the random variable (move the risk adjusted rate of return to the risk free, which is lower by assumed risk aversion). It is clearly seen that for high values of V the mass associated is lower under the risk neutral distribution, hence if the real distribution is used to value option it would be overvaluing its true value. This insight confirms our previous derivations. In the same

[22] Otherwise they will not add up to one.

Figure 1. Change in Drift and Redistribution of Density Mass for a Positive Risk Premium on a Normal Distribution



tense, for a low value of V the mass associated is lower, but this change does not affect the value of the option, which has positive value only for high realizations of V (otherwise is zero, never negative).

Remark: If the risk premium is negative, as would be the case if under the CAPM world the underlying asset happens to have a negative covariance with the market return, and hence a negative premium, the problem arising will be of undervaluation.

IV. Results

Due to result obtained, though the valuation for the underlying asset is the same under both mechanisms, when it comes to evaluate growing cash flows (horizontal, vertical or within the same market) embedded in the project, the traditional DCF overvalues the true option value. Although the discounting rate is smaller (and hence the discount coefficient is greater, which leads to increase the value of the option calculated by risk neutral valuation) this effect cannot offset the decrease in expected value due to the application of the new probability distribution.

As it was shown, the use of the true distribution and a risk adjusted rate when the applicable distribution is the risk neutral (or synthetic) with the risk free rate lead to overvaluation due to the asymmetry of the payoffs. Consider for instance a start up project. If for valuation purposes we forecast growing cash flows and a residual value consistent with them, and growth has to be supported by periodical investments until it reaches its mature stage, the value thus obtained will imply exercising successive growth options. Given that the value at the mature stage includes exercised growth options, there would be a tendency to overstate the true value of the start up. The degree of overvaluation will depend upon the values adopted by the following parameters: r (risk free rate), k (risk adjusted rate), p (probability of high values for the project), V_u (the value of the project in a good state) and V_d (the value of the project if things do not go too well).

A. Comparative Statics

A simulation model can provide more insights. Assume the two possible values the company can take are 135 in one scenario (with probability 43%) and 95 in the other (with probability 57%). The risk-adjusted discount rate is assumed to be 10%. Under the traditional DCF methodology, the value of the project would be 100. Now assume that at the following period the company is able to expand further by paying a cost of 200 to obtain an expected value of two times the value of the company at t + 1. This growth opportunity will be exercised only if the market proves to be good for the company (scenario 1). For the purposes of comparative static we change one parameter at a time, keeping the others constant. In Table 1 we can observe the results of our changes in the values of the parameters;[23] V_u is the value of the company in the good state of the nature, V_d in the bad state, r is the risk free rate, k is the risk adjusted discount rate, p is the true probability of the good state of the nature, and the expansion payoff (growth in cash flows) is the function *Max* $(2\,V_i - K)$.

We first change the upper value of V, then the lower value of V, we continue by changing the risk free rate and the risk adjusted rate of return, and finally we change the value of the true probability p. The results are the following:

[23] The results are based upon movements of Vu to 140, Vd to 85, r to 7%, k to 12% and p to 50%. In the last row the degree of arising overvaluation can be seen.

Table 1. Simulation Parameters and Results for Comparative Statics

	Initial value	Vu =140	Vd = 85	r = 7%	k = 12%	p = 50%
Present value of the asset	100	103.9	94.8	100	98.2	102.3
Risk neutral probability (p)	29%	31%	32%	34%	23%	35%
Growth option value under DCF	23.4	31.2	23.4	23.4	23.0	27.3
Growth opt. val. under risk neutral valuation	16.3	23.9	18.5	19.2	13.3	20.2
Extent of overvaluation	44%	31%	27%	22%	73%	35%

Notes: The initial values for the parameters are the following: upside value, V_u = 130; downside value, V_d = 95; riskfree rate, r = 5%; discount rate, k = 10%; probability of upside scenario, p = 43%; expansion payoff, 2 times the current value V_i = 2 V_i; cost of investment of expansion, K = 200; and net payoff of expansion, Max (2 V_i - K, 0) = 260.

(a) an increase on the upper possible value V_u reduces the excess of overvaluation, (b) a decrease on the lower possible value V_d reduces the extent of overvaluation, (c) an increase on the risk free rate r reduces the excess of overvaluation, (d) an increase on the risk adjusted discount rate k increases the excess of overvaluation, and finally, (e) an increase on the real probability p of upward movements reduces the degree of overvaluation.

Now we shall explain the intuition underlying these effects from the formula for calculating risk neutral probabilities in our simple model; the probability $\tilde{p}$ comes from[24] the following formula:

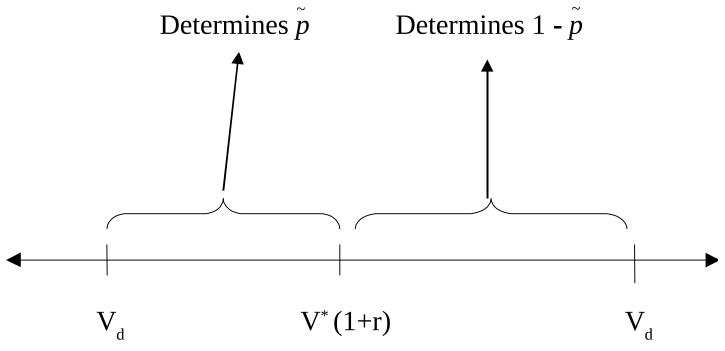
$$\tilde{p} = \frac{V\,(1+r) - V_2}{V_1 - V_2}$$

[24] See Appendix 2.

This can be better appreciated with the help of Figure 2, where it can be seen how the value V_u and V_d, together with the initial value V and the risk free rate r give rise to the risk neutral probability in a binomial world. An increase on the upper value V_u increases the expected value of the underlying asset. Given the methodology of calculation of the risk neutral probability $\tilde{p}$, we would expect the probability to diminish, however, this effect is more than offset by the move in the expected value of the asset (used together with the risk free rate of return to determine the risk neutral probabilities), which moves the division line between probabilities to the right. This effect overcomes the other, hence increasing $\tilde{p}$. This situation drives the risk neutral probability closer to its real counterpart (which is assumed to be constant here), reducing the extent of overvaluation.The decrease on V_d leads to the same effect. The changes on this extreme value are exactly the opposite as those described previously (the upper value going up is equivalent to the lower going down). In both cases the expected value of the underlying asset is affected, though in the opposite sense, impacting on the divisory line between risk neutral probabilities. An increase on V_u or a decrease on V_d broadens the range between the extreme values, affecting in an opposite way the expected value of the underlying asset but affecting in the same way the risk neutral probability, bringing it closer to the real counterpart, therefore reducing the degree of overvaluation.

Both an upward movement on the risk free rate r, or a reduction on the risk

Figure 2. Determination of the Risk Neutral Value for p from the Parameters of the Simulation



adjusted rate k, can be synthesized in a change on the risk premium of the asset (the risk adjusted rate can be decomposed into two components, the risk free component and the risk premium).

An increase of r (keeping k constant) as well as a decrease on k (given r), can be assimilated to a decrease on the equilibrium risk premium. However, the effects on the dependent values are not exactly the same.[25] An increase of r does not change the expected value of the asset, but affects the line dividing the risk neutral probabilities. Given how this probability $\tilde{p}$ is calculated, the division line is moved to the right, increasing it. This drives the risk neutral probability closer to the real probability, therefore reducing the extent of overvaluation.

The effect of an increase of k affects the expected value of the underlying asset moving the division line to the left, thereby reducing the risk neutral probability $\tilde{p}$ and broadening the gap between the synthetic and the real probability.

Finally, an increase of p increases the expected value of the underlying asset. This moves the division line to the right, therefore increasing $\tilde{p}$ and reducing the degree of overvaluation.

It stems from these explanations that the analysis mainly passes through the study of the movements of the division line that makes up the values of the risk neutral probabilities $\tilde{p}$ y 1-$\tilde{p}$. It is not complicated to find out from a visual inspection the consequences of movements on the value of the parameters.

B. An Application

In recent works[26] a methodology has been suggested to value internet[27] and technology companies (and by extension applicable to any start up project). This methodology is also used in the "venture capital valuation"

[25] In fact the effects are the opposite.

[26] See Desmet, Francis, Hu, Koller, and Riedel (2000).

[27] The case study used is Amazon.com; roughly speaking, it is calculated the expected value of Amazon in 2010, using estimates of market share in different segments of business. The value is then discounted by means of a risk adjusted rate to the present to obtain the current value.

model.[28] The method works backwards, starting by obtaining the would-be value, which can be thought as the expected value, of the company at some point in the future, when it is consolidated and making profits. This expected value is then discounted using a risk-adjusted discount rate to obtain the value of the company today, being this methodology consistent with (1) and (4). Here our analysis starts to be applied; consider the value of the company in the future, in some years time; this value is reached after several investments outlays are made. Each of the installments is contingent on previous growth attained, so as long as nature shows up favorable for the project, new investment takes place to keep the growth rate. We are able then to say that the value in the future is contingent on nature showing favorable[29] until it reaches such a point.

If we then value the business by DCF, we would be falling into the overvaluation problem previously described. Our analysis suggests that by valuing contingent (on growth) streams of cash flows[30] using the discounted cash flow methodology, the value of the business will tend to be overestimated. The situation previously described is shown in Figure 3, where it is clearly seen that there is a reinvestment pattern (which is contingent on previous events) needed to attain the growth of cash flows and the value of the project at the mature stage. By directly discounting growing cash flows and residual value (methodology widely used to value high risk long-term projects, like the ones we deal with) will be falling under the problem described.

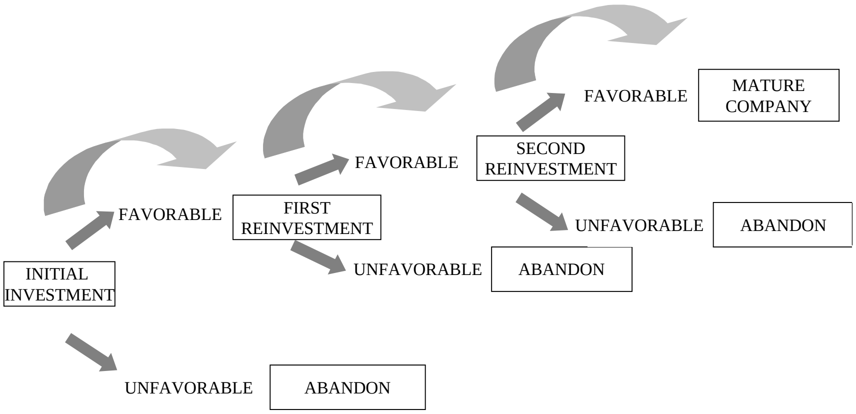
To the purpose of solving the problem of overvaluation detected and exposed previously, the following methodology is proposed to correctly evaluate the growth opportunities: (a) separate the outcomes of contingent decisions from the current value of the company, (b) analyze the random structure of events the company faces, (c) define a variation range for the possible values of the business, without including results of options, (d) calculate the present value using the DCF method, to determine the value of the underlying asset, and with this in hand, determine the risk neutral

[28] See Sahlman and Scherlis (1989).

[29] With the help of the management as well.

[30] We are able now to see how important were contingent payoffs just by taking a look ate the current economic and financial situation of the company.

Figure 3. Contingent Investment Sequence Needed to Maintain the Pattern of Growth for High Growth Companies



probability distribution, (e) use these probabilities to value the options, discounting the expected value to the risk free rate, and (f) add the value thus determined to the value of the company.

We know it is not an easy task, and that we have worked with a simplified model. However, the fact of thinking about contingent situations and possible outcomes represents a great advance to the company and manager's strategic thinking.

V. Conclusions

A now growing literature on real options is taking advantage of the theory and practice of financial options. It starts to be thought that options are everywhere within the company, and given that flexibility has value, the real option framework is the appropriate method to capture it. Throughout this paper it has been demonstrated that growth patterns in cash flows of high growth companies or projects embed growth options through successive investments and reinvestments, which if valued using through straight traditional DCF may give rise to an overvaluation problem.

The intention of this paper was to show that valuation of projects and business with growth opportunities must take into account the overvaluation effect they are exposed to, given that future value is contingent on favorable events. The present value of a business is composed of two elements: the present value of assets in place and the growth opportunities.

The weight of each component will be affected by the industry and the firm's own characteristics. To the extent that the company is in a mature industry, and the possibility of growing has been fully exploited and reflected in the current value of the firm assets, the growth component will tend to be relatively not significant with respect to the full value, so reinvestment needs will not be significant. On the other hand, for companies and industries in expansion or in newly created industries, the most of the value will be captured by growth options due to the need of reinvesting heavily, weighing more significantly in the full value. This contingent growth will have associated a high volatility, due primarily to the uncertainty surrounding the market, the product or service, competitors and substitutes. Being more significant the option component for this kind of industries, the use of the traditional DCF model for valuation purposes will offer more problems, prompting overvaluation.

The most significative and illustrative example can be captured by the impact of technology and globalization on growth opportunities of companies and industries. This affects industries asymmetrically and to different extents. For those companies that are affected the most, technology creates a complete new world of opportunities, and also creates risk of overvaluing business due to the problems described, under the assumption that investors use the DCF model as a valuation tool. Options must be valued as their nature claims.

However, it has been shown throughout this paper that both methods are complements rather than substitutes. Risk neutral probabilities cannot be obtained without figuring out the current value of the underlying asset, for which DCF is appropriate; so they work together towards the same goal. Nevertheless, each method has to be applied for the right situation to a proper analysis of the allocation of resources.

Our results are derived based upon a set of assumptions, so results are conditioned and the model developed is not very complicated. However, these assumptions are not more restrictive than those involved in the derivations of models like the Capital Asset Pricing Model or the Black Scholes formula. Nevertheless, this fact should not stop us from relaxing assumptions and searching for new results. This is a very attractive topic for future research.

Appendix 1

It follows that at t + 1 an asset with payoffs of V_i in each state of the nature i is worth,

$V_t = \Sigma\, \pi_i\, V_{i,t+1}$ at t.

Working on this formula, multiplying and dividing by $\Sigma\ \pi_i$ and redistributing, we obtain,

$$V_t \;=\; \Sigma \pi_i V_{i,\ t+1} \frac{\Sigma \pi_i}{\Sigma \pi_i} \;=\; \Sigma \frac{\pi_i}{\Sigma \pi_i}\ V_{i,\ t+1}\ \Sigma \pi_i$$

and making, $\tilde{p}_i = \frac{\pi_i}{\Sigma \pi_i}$, and $\Sigma \pi_i = \frac{1}{(1+r)}$, we obtain

$$V_t \;=\; \Sigma\, \tilde{p}_i\ V_{i,\ t+1}\ \frac{1}{(1+r)}$$

Appendix 2

In short, the changes introduced are: (a) take the current value of the asset, (b) set its return equal to the risk free return, (c) find the probabilities associated to this new expected value by changing the probability mass at each point of the possible values of V. In formula,

$$V \;=\; [\tilde{p}\ \ V_1 \;+\; (1-\tilde{p})\ V_2]\ \frac{1}{(1+r)}$$ rearranging terms,

$$(1+r)\ V \;=\; \tilde{p}\ \ V_1 \;+\; (1-\tilde{p})\ V_2$$ can be easily solved for

$$\tilde{p} \;=\; \frac{V\ (1+r)\ -\ V_2}{V_1\ -\ V_2}$$ and $$(1-\tilde{p}) \;=\; \frac{V_1\ -\ (1+r)\ V}{V_1\ -\ V_2}$$

References

Black, F., and M. Scholes (1973), "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy* **81**: 637-659.

Brealey, R. and S. Myers (1996), *Principles of Corporate Finance*, Mc Graw Hill, fifth ed.

Constantinides, G. (1978), "Market Risk Adjustment in Project Evaluation," *Journal of Finance* **33** (2): 603-616.

Cox, J., Ross, S., and M. Rubinstein (1979), "Option Pricing: A simplified Approach," *Journal of Financial Economics* **7** (3): 229-263

Desmet, D, Francis T., Hu, A., Koller, T., and G. Riedel (2000), "Valuing dot coms," *McKinsey Quarterly Journal* **1**: 150-157.

Dixit, A., and R. Pindyck R. (1994), *Investment under Uncertainty*, Princeton University Press, Princeton, N.J.

Hull J. (1997), Options, *Futures and other Derivative Securities*, Prentice Hall, third ed.

Kester, W.C. (1984), "Today's Options for Tomorrow Growth," *Harvard Business Review* **62** (2): 153-160.

Kester, W.C. (1993), "Turning Growth Options into Real Assets," in *Capital Budgeting under Uncertainty*, R. Aggarwal, ed., Prentice Hall.

Kulatilaka, N. (1995), "The Value of Flexibility: A Model of Real Options," in L. Trigeorgis, ed., *Real Options in Capital Investment*, Praeger.

Kulatilaka, N. and A. Marcus (1992), "Project Valuation under Uncertainty: When does DCF Fail?," *Journal of Applied Corporate Finance* **5** (3): 92-100.

Mason, S.P., and R.C. Merton (1985), "The Role of Contingent Claim Analysis in Corporate Finance," in E. Altman and Subrahmanyam, eds., *Recent Advances in Corporate Finance*, Irwin.

Mendenhall, W., Beaver, R., and B. Beaver (1998), *Introduction to Probability and Statistics*, Duxbury.

Merton, R.C. (1973), "Theory of Rational Option Pricing," *Bell Journal of Economics and Management Science* **4** (1): 141-183.

Myers, S. (1977), "Determinants of Corporate Borrowing," *Journal of Financial Economics* **5**.

Neftci, S. (1996), *An Introduction to the Mathematics of Financial Derivatives*, Academic Press.

Ross, S. (1998), *Corporate Finance*, McGraw-Hill.

Sahlman, W., and D. Scherlis (1989), “A Method for Valuing High-risk Long-term Investments,” *Harvard Business School*, 9-288-006.

Trigeorgis, L. (1997), *Real Options: Managerial Flexibility and Strategy in Resource Allocation*, The MIT Press, Cambridge Massachussets.

Trigeorgis, L. (1988), “A Conceptual Options Framework for Capital Budgeting,” *Advances in Futures and Options Research* **3**:145-167.

Varian, H. (1992), *Microeconomics Analysis*, WW Norton & Company.

Willner, R. (1995), “Valuing Start-up Venture Growth Options,” in L. Trigeorgis, ed., *Real Options in Capital Investment*, Praeger.

Journal of Applied Economics, Vol. VI, No. 1 (May 2003), 73-93

MEASURING COMPETITION IN THE U.S. AIRLINE INDUSTRY USING THE ROSSE-PANZAR TEST AND CROSS-SECTIONAL REGRESSION ANALYSES

Thorsten Fischer[*]
Economy.com, Inc.
and
David R. Kamerschen[*]
The University of Georgia

Submitted August 2002; accepted February 2003

We employ the Rosse-Panzar test to assess market performance in selected airport-pairs originating from Atlanta. The Rosse-Panzar test stands in the tradition of the New Empirical Industrial Organization. It is based on the comparative statics of a reduced form revenue equation. Therefore, it is less powerful than structural models, but it offers the advantage of less stringent data requirements and reduces the risk of model misspecifications. The test statistic allows us in most airport-pairs to reject both conducts consistent with the Bertrand outcome, which is equivalent to perfect competition, and the collusive outcome, which is equivalent to joint profit-maximization. Rather, the test statistic suggests that behavior is consistent with a range of intermediate outcomes between the two extremes, including, but not limited to the Cournot oligopoly. In the second part of the paper, a cross-section pricing regression complements the Rosse-Panzar test. It shows that the presence of low-cost competition in an airport-pair reduces the average fare significantly.

JEL classification codes: L00, L40, L93

Key words: airlines, oligopoly, conduct, price-cost margins, Lerner index, Rosse-Panzar test

[*] The authors are Economist, Economy.com, Inc., and Distinguished Professor of Economics and Jasper N. Dorsey Chairholder, Terry College of Business, The University of Georgia, respectively. We are heavily indebted to Drs. Scott E. Atkinson, Charles D. DeLorme, Jr., Claudia E. Halabi, Donald C. Keenan, Peter G. Klein, David B. Robinson, and an anonymous referee for their ideas and suggestions.

I. Introduction

The U.S. airline industry has experienced revolutionary change in the last two decades moving from strict regulation to modest regulation, now allowing airlines to decide such things as their pricing strategies, frequency of schedule, and entry into and exit from markets. However, access to some key inputs, such as airport boarding sites, is still determined by non-market or regulatory conditions. Proponents of deregulation expected better performance through enhanced competition, resulting in higher productivity, lower costs, lower fares, and better service. This optimism has been largely fulfilled as the U.S. airline industry in recent years has had steady growth, falling prices, more convenient schedules, and moderate concentration, although profits have been rather volatile (see, e.g., Bailey, 2002, Gowrisankaran, 2002). It can be argued that since the late 1980s and early 1990s, the industry has settled into a new equilibrium. The vital and challenging question is whether this (less than ideal) deregulated market performed better than before, or whether there still exists market power and market conduct that are less optimal than previously.

This paper examines the economics underlying the U.S. airline industry, and its development and evolution since deregulation. More specifically, this paper studies the pricing strategy, market conduct, and market performance in the U.S. airline industry in recent years. Two empirical models are employed, each with a different focus and methodology. The level of analysis is on the micro-level, concentrating on the firm and airport-pair level. This enables a more detailed and precise approach to the study of market conduct than would be feasible with more aggregated data.

The statistical analysis is restricted to airport-pairs originating in Atlanta. Atlanta is an appropriate choice for conducting such a study for several reasons. First, Atlanta serves as a major hub for Delta Air Lines, one of the nation's largest carriers. Delta accounts for more than eighty percent of all departures and arrivals at Atlanta's Hartsfield International airport. Therefore, any effects that a dominant firm may have on the market's competitiveness are captured. Second, Atlanta is an important market for all other major U.S. carriers that compete with Delta by offering one-stop service to most cities connecting through their respective hubs. Third, Atlanta has experienced entry by a remarkably successful lowcost carrier, ValuJet Airlines, which started in 1993 and grew rapidly. At its peak, it served almost 30 markets and used more than 50 aircraft. After the loss of one of its planes in May 1996, ValuJet was grounded for approximately three months and is still struggling to rebuild its former position. ValuJet faced severe restrictions imposed by regulators on its growth opportunities. Furthermore, consumer confidence in its safety and reliability suffered immensely. In July 1997, Valujet Inc., the parent of ValuJet Airlines, announced plans to merge with Florida-based Airways Corp., parent of AirTran Airways. The merger took effect with the larger carrier, ValuJet, adopting the smaller carrier's name, AirTran, to eliminate any association with the crash. The Orlando-based AirTran Airways with its hub in Atlanta has experienced steady growth and consolidated its position as a successful provider of lowcost air travel. Early in 2000, it took delivery of the first of 50 new-generation Boeing 717 aircraft, in pursuit of its strategy of growth and modernization of its fleet. In 2002, AirTran was named Airline of the Year for the fourth consecutive year by the American Society of Travel Agents. The trade group said it honored the discount airline for creating an Internet booking engine aimed at travel agents, and for continuing to provide competition in the industry. Most big carriers, including Delta, eliminated base travel agent commissions in 2002. Of course, what long-run impact the terrorist attacks on the U.S. on September 11, 2001, will have on AirTran and indeed the entire U.S. airline industry is hard to predict at this time.

Our format provides an interesting opportunity to study market conduct in different competitive environments: markets where Delta is the only carrier, markets where Delta competes with other majors, and finally markets where Delta competes against a lowcost, start-up carrier. Anecdotal evidence suggests that after the grounding of ValuJet, airfares in certain markets rose sharply. One well-publicized example is the route linking Atlanta and Mobile, AL, where the coach fare rose from $79 to more than $400. Some communities in the Southeast provided financial incentives to ValuJet to induce the carrier to serve their airports.

New Empirical Industrial Organization (NEIO) research identifies and estimates the degree of market power, specifies and estimates the behavioral equations that drive price and quantity, and often infers marginal cost or measures market power without it. NEIO studies emphasize individual industries, because institutional details make broad cross-section studies of industries of limited value. NEIO provides techniques to execute studies on market conduct and market power of individual industries by estimating empirically parameters of conduct that identify well-defined models of oligopoly. The estimated values in conduct studies such as this one cover the range of distinct behavior from the Bertrand case on one end, through the Cournot oligopoly, to the collusive cartel outcome on the other end. Thus, the estimates thus provide a numerical equivalent to oligopoly conduct ranging from perfect competition to joint profit-maximizing monopoly.

Structural models, based on oligopoly theory, can be tailored to the idiosyncrasies of the particular market under investigation, obviating restrictive assumptions about symmetry across industries. Moreover, the degree of market power is directly estimated from the data. This permits explicit hypothesis testing of the degree of market power and industry conduct. Where structural models are not feasible because the relevant data are not available, or the validity of the specified structural model is in question, reduced-form approaches are useful to distinguish firm conduct and market power. These reduced-form approaches are generally less powerful than structural models, but they impose less demanding data requirements, and reduce the risk of employing an ill-specified model. Reduced-form approaches are often non-parametric, and rely on the comparative statics of some economically relevant function.

This paper investigates market conduct and performance by employing a non-structural model in the tradition of the NEIO. The so-called Rosse-Panzar test is based on the reduced revenue function of the firm and determines market structure by analyzing comparative statics of the total revenue function with respect to cost. The study uses firm-level data aggregated from raw balance-sheet data, employing index number theory, thereby obtaining very accurate measures of input prices. An improved approach is used to compute the price of capital. The Capital Asset Pricing Model (CAPM) is employed to obtain a reasonably accurate measure of the opportunity cost of capital. This measure is superior to conventional measures that rely on accounting rather than economic concepts of capital pricing. The paper also employs airport-pair-level data on airfares, thus allowing a degree of detail that renders the study very valuable for investigators interested in specific competitive set-ups rather than a broader and more general framework. The sample extends over the 24 quarters from January 1991 to December 1996. Finally, a cross-section regression model is employed to supplement the studies on market structure, to provide additional insight into pricing strategies, and to explore the factors that influence the price of air travel.

Section II presents an approach to testing for monopoly behavior, the Rosse-Panzar test, which allows for a first impression regarding market conduct. Section III implements the Rosse-Panzar test empirically and presents the results. Section IV presents a cross-section regression for the Atlanta market to assess the impact of a lowcost carrier on fares. Section V briefly concludes with the major findings.

II. Theoretical Background

Rosse and Panzar (1977) and Panzar and Rosse (1987) introduce a series of tests based on properties of reduced-form revenue equations at the firm level on which the hypothesis of monopoly or oligopoly profit maximization places testable restrictions.[1] The data requirements, consisting of revenues and factor prices, are relatively modest. The following model is taken from Panzar and Rosse (1987) and follows their development of the test closely.

Let q be a vector of decision variables that affect a firm's revenue. In the most natural case q would describe a vector of output quantities. Let z denote a vector of variables that are exogenous to the firm and shift the firm's revenue function. The firm's cost function also depends on q, so that C = C (q, w, t), where w is a vector of factor prices also taken as given by the firm and t is a vector of exogenous variables that shift the firm's cost curve.[2] It follows that the firm's profit function is given by

$$\pi = R - C = \pi(q, \quad z, \quad w, \quad t) \tag{1}$$

Let q^0 be the argument that maximizes this profit function. Also, let q^1 be

[1] For an extension of the Rosse and Panzar test when variables besides the firms' revenues are observable, see Sullivan (1985) and Ashenfelter and Sullivan (1987).

[2] While this cost function ignores efficiencies generated by hubs, these cost complementaries do not make the Rosse-Panzar result inapplicable.

the output quantity that maximizes π (q, z, (1 + h) w, t) where the scalar h is greater or equal to zero. Define R^o as R (q^o, z) ≡ R^* (z, w, t) and R^1 = R (q^1, z) ≡ R^* (z, (1 + h) w, t), where R^* is the firm's reduced form revenue function. It follows by definition that

$$R^1 - C(q^1,(1+h)\ w,\ t)\ \geq R^0 - C(q^0,(1+h)\ w,\ t) \tag{2}$$

Using the fact that the cost function is linearly homogeneous in w, this can be written as

$$R^1 -\ (1+h)\ C(q^1,w,\ t)\ \geq R^0 -\ (1+h)\ C(q^0,w,t) \tag{3}$$

and that

$$(R^1 - R^0)/h\ =\ [R^*(z,\ (1+h)\ w,\ t)\ -\ R^*(z,\ w,\ t)/h]\ \leq\ 0 \tag{4}$$

This is the non-parametric result that indicates that a proportional cost increase will result in a decrease of the firm's revenues. Assuming that the reduced-form revenue equation is differentiable, taking the limit of (4) for h → 0 and dividing by R^* yields

$$\Psi^* \equiv\ \Sigma\ w_i\ (\delta R^*/\delta w_i)/R^* \leq\ 0 \tag{5}$$

where the w_i are the components of the vector w, so that w_i denotes the price of the ith input factor.

This describes a restriction imposed on a profit-maximizing monopoly. The sum of the factor price elasticities of the reduced-form revenue equation cannot be positive. Intuitively, the question that the test statistic ψ^* tries to answer is what is the percentage change in the firm's equilibrium revenue resulting from a one-percent increase in all factor prices. An increase in factor prices shifts all cost curves, including the marginal cost curve, up. Consequently, the price charged by the monopolist goes up and the quantity decreases. Since the monopolist operates on the elastic portion of the demand curve, total revenue decreases. Hence, ψ^* is non-positive. The generality of the result causes one drawback for the test. Even for "monopolies" facing a perfectly elastic demand curve, the value for ψ^* is less than zero. All firms which operate in isolation, that is, all firms whose structural revenue functions do not depend on any other agent's decisions, will show a test statistic that is non-positive. Therefore, a rejection of the hypothesis that ψ^* is less than zero must indicate that the firm is affected by other agents' actions.

The next question, then, is whether there exist any models consistent with an estimate for ψ greater than zero. Fortunately, this is the case. Rosse and Panzar cite three models of equilibrium consistent with a positive value for ψ. In all three models, the revenue function facing the firm depends on the action of potential or actual rivals. In other words, the firm no longer acts in isolation. The results for the models depend crucially on the assumption that the observed firms be in long-run equilibrium. We will restrict our attention to two additional models that are interesting with respect to airlines. First, the benchmark case of the long-run competitive equilibrium is examined, and subsequently the conjectural variation oligopoly is explored. Unless some kind of interaction between firms is introduced into the model dealing with perfect competition, price-taking behavior will lead to a ψ^* less than zero. The output price that a firm faces, therefore, is endogenized by allowing for competitive entry and exit. This model has been discussed most prominently by Silberberg (1974). The reasoning is as follows. Changes in factor prices will, at least in the longrun, lead to exit or entry and consequently to changes in output prices. These changes in turn will affect input demand and output supply decisions of the firm.

For firms observed in long-run equilibrium, the sum of the elasticities of reduced form revenues with respect to factor prices equals unity (Rosse and Panzar, 1987). The intuition behind this result is that a one-percent increase in all factor prices will result in an equal-proportional that is one-percent, increase in total revenue. Because average cost is homogeneous of degree one in w, a one-percent increase in all factor prices will shift the average cost curve up by one percent for all output levels. Consequently, the minimum point is unchanged. Since in long-run competitive equilibrium the firm operates at minimum average cost, the competitive output q^c remains unchanged. However, in equilibrium, the competitive price p^c must be equal to minimum average cost, which has increased by one percent. Therefore, p^c must have increased by one per cent also, driving up total revenues by the same percentage. Therefore the condition that ψ^c be equal to one is established.

Contrast this with the result if firms are not in long-run equilibrium. More specifically, assume we observe a firm after the one-percent increase in all factor prices, but before any firms have exited from the market. The firm will respond by reducing output while the price remains initially unchanged, thus resulting in a decrease in total revenues. Hence, in the shortrun, ψ is less or equal to zero. Only after some firms exit does the price go up to the new long-run equilibrium level and is output restored to its original level. This should underline the importance of the long-run equilibrium assumption.

The final point to be made is that a conjectural variations oligopoly model that exhibits strategic interactions among a fixed number of rivals may also be consistent with positive values of ψ. Only if the oligopoly behaves close to a joint monopoly, that is, if firms collude, is the marginal industry revenue positive.

In summary, we have provided a non-structural test for the existence of monopoly power, and we have derived three important results.[3] First, the sum of elasticities of revenue with respect to each input price is negative in monopoly or collusive (joint monopoly) equilibrium. It is also negative in short-run competitive equilibrium. Moreover, it is equal to unity in long-run competitive equilibrium and indeterminate in a general conjectural variation oligopoly equilibrium. These implications can be tested empirically. For instance, a finding of a test statistic Ψ that is positive, would rule out monopoly or a collusive cartel equilibrium.

A profit-maximizing monopolist operating on the elastic portion ($\eta < -1$) will exhibit a negative value for Ψ. It also demonstrates that a negative sign cannot rule out competition since a competitive firm tends to face an even more elastic demand curve. Using the result obtained previously, Shaffer (1982a), Shaffer (1983a) derives the Lerner index (L_j) in terms of the Rosse-Panzar test statistic where s_j is firm j's market share.

[3] While the focus of empirical IO has shifted away from identifying conjectures parameters in simply quantity-setting models to identifying demand and costs in differentiated price-setting models, we think the conjectures equilibrium framework with quantity competition and the cross-sectional regressions are still a useful methodology. To see the newer focus, see, e.g., Berry's 1992 paper on airline competition where he estimates a model of customer heterogeneity (business vs. leisure) which is important in this industry because of price discrimination.

We obtain the Lerner index for an individual firm and for the industry as a whole, respectively.

$$L_j = 1/(1 - \Psi_j) \tag{6}$$

and

$$L = (H + \Sigma\ s_j^2\, \lambda_j)\ \ /\ \left[s_j(1 + \lambda_j)\ (1 - \Psi_j\right] \tag{7}$$

Equations (6) and (7) express the firm and industry Lerner indices, respectively, as a function of market share, the conjectural variation parameter λ and the Rosse-Panzar test statistic *H*. The firm's Lerner index depends only on the test statistic, which is independent of market share or the conduct parameter. The result is valid only as long as the short-run equilibrium is considered, that is, changes in total revenue due to changes in factor prices before entry and exit occur. In a further paper, Shaffer (1983b) extends his result found in 1982 to a more general connection between the Rosse-Panzar statistic and the price elasticity of demand.

The reduced-form revenue equation has been used as a test of market power among others by Shaffer (1982b), Nathan and Neave (1989), and Shaffer and DiSalvo (1994). In all cases, the test has been applied to the banking industry. Furthermore, Shaffer and DiSalvo apply both tests, i.e. the conjectural variations oligopoly and the Rosse-Panzar test, to a duopoly banking market in Pennsylvania. This is a procedure we follow.

III. Empirical Strategy

A. Implementation of the Rosse-Panzar Test

To apply the Rosse-Panzar test, we need to derive a reduced-form revenue equation. However, we must also consider the underlying structural model in developing the reduced form. Following Shaffer and DiSalvo, we propose the estimation of the following equation, taking into account that output quantity is endogenous. The demand equation is given by (8), and a total revenue equation is added in loglinear form. Alternatively, the translog specification could be used. The loglinear revenue equation is given as

$$\ln\ TR = b_0 + b_1 \ln q_j + \Sigma c_i \ln w_{ij} \qquad (8)$$

where i = 1,..., 4 denotes inputs and the subscript j denotes airlines. *TR* denotes total revenue, *q* denotes output and *w* denotes factor prices. The parameters to be estimated are b_0, b_1 and c_1 through c_i.

The equations are estimated separately for each carrier using a generalized methods of moments approach. We employ price and quantity data for outbound traffic, year dummies and their interaction term as instruments for inbound traffic, and inbound data as instruments for outbound data. The instruments make for a very good fit, since they are highly correlated with the right-hand variables and almost uncorrelated with the error term. It is clear from equation (8) that the sum of the estimates for c_i yields the required test statistic Ψ.

Table 1. Estimates of the Rosse-Panzar Test Statistic for Outbound Traffic, Ranked from Lowest to Highest

	City - Pair	RP-Statistic (outbound)	Standard errors*
1	Washington Dulles Intl. - United (IAD-UA)	-20.2920	2.56568
2	Miami Intl. - American (MIA-AA)	-6.03789	2.45715
3	Philadelphia Intl. - Delta (PHL-DL)	-5.70006	2.49875
4	Memphis Intl. - Delta (MEM-DL)	-5.51766	1.44122
5	Chicago O'Hare Intl. - American (ORD-AA)	-4.79100	2.20212
6	Miami Intl. - Delta (MIA-DL)	-4.50376	1.82038
7	Chicago O'Hare Intl. - Delta (ORD-DL)	-3.98305	1.18051
8	George Bush Intl. Continental/Houston - Delta (IAH-DL)	-1.89520	1.22414
9	Detroit Metropolitan Wayne County Intl. - Delta (DTW-DL)	0.071092	1.82740
10	Newark Intl. - Delta (EWR-DL)	1.87128	2.22949
11	Boston Intl. - Delta (BOS-DL)	2.7669	2.65899
12	Lambert St Louis Intl. - Delta (STL-DL)	3.80627	1.66572
13	Pittsburgh Intl. - US Air (PIT-US)	3.89118	2.51419

Table 1. (Continued) Estimates of the Rosse-Panzar Test Statistic for Outbound Traffic, Ranked from Lowest to Highest

City - Pair	RP-Statistic (outbound)	Standard errors*
14 Minneapolis St Paul Intl/Wold-Chamb. - Delta (MSP-DL)	4.07826	1.18344
15 Washington Dulles Intl. - Delta (IAD-DL)	4.59163	2.10702
16 Pittsburgh Intl. - Delta (PIT-DL)	4.67703	1.22940
17 Memphis Intl. - Northwest (MEM-NW)	4.81010	1.28097
18 La Guardia - Delta (LGA-DL)	7.55154	1.70849
19 Ronald Reagan Washington Natl. - Delta (DCA-DL)	7.69299	1.07561
20 Philadelphia Intl. - US Air (PHL-US)	9.73294	2.72694
21 Detroit Metrop.Wayne County Intl-Northwest (DTW-NW)	10.6878	1.55307
22 Newark Int. - Continental (EWR-CO)	10.7625	4.21756
23 Charlotte Intl. - Delta (CLT-DL)	12.2199	2.64956
24 G. Bush Intl. Continental/Houston - Continental (IAH-CO)	13.0914	2.09673
25 Dallas Ft. Worth - American (DFW-AA)	13.3728	1.73529
26 Minneapolis St Paul/Wold-Chamb.- Northwest (MSP-NW)	13.6637	2.73482
27 Charlotte Intl. - US Air (CLT-US)	15.1083	3.85622
28 Chicago O'Hare Intl. - United (ORD-UA)	16.8336	4.38717
29 Dallas Ft. Worth - Delta (DFW-DL)	17.1838	2.86400

Note: *All coefficients have a significantly positive test statistic, which is also significantly different from one.

Table 2. Estimates of the Rosse-Panzar Test Statistic for Inbound Traffic, Ranked from Lowest to Highest

	City - Pair	RP-Statistic (inbound)	Standard errors*
1	Washington Dulles Intl. - United (IAD-UA)	-23.999	4.62528
2	Philadelphia Intl. - Delta (PHL-DL)	-7.12364	2.68764
3	Miami Intl. - American (MIA-AA)	-4.73940	2.63880
4	Memphis Intl. - Delta (MEM-DL)	-4.15026	1.91954
5	Chicago O'Hare Intl. - Delta (ORD-DL)	-4.14051	1.04254
6	George Bush Intl. Continental/Houston - Delta (IAH-DL)	-3.94652	1.27333
7	Chicago O'Hare Intl. - American (ORD-AA)	-3.73036	1.67664
8	Miami Intl. - Delta (MIA-DL)	-3.48789	2.16008
9	Detroit Metropolitan Wayne County Intl. - Delta (DTW-DL)	-0.465805	1.75467
10	Pittsburgh Intl - US Air (PIT-US)	-0.262022	2.47935
11	Charlotte Intl.- Delta (CLT-DL)	0.000039	0.000013
12	Charlotte Intl. - US Air (CLT-US)	0.00032	0.000013
13	Pittsburgh Intl. - Delta (PIT-DL)	1.47711	1.68889
14	Newark Intl. - Delta (EWR-DL)	2.21861	2.16485
15	Boston Intl. - Delta (BOS-DL)	2.51153	2.15238
16	Lambert St Louis Intl. - Delta (STL-DL)	3.78565	1.67995
17	Minneapolis St Paul Intl/Wold-Chamb. - Delta (MSP-DL)	3.80256	1.27049
18	Memphis Intl. - Northwest (MEM-NW)	4.81165	1.48858
19	Washington Dulles Intl. - Delta (IAD-DL)	5.80216	2.18200
20	La Guardia - Delta (LGA-DL)	6.25637	1.30126
21	Ronald Reagan Washington Natl. - Delta (DCA-DL)	6.64213	1.10359
22	Detroit Metrop. Wayne County Intl-Northwest (DTW-NW)	8.63238	1.34562
23	Philadelphia Intl. - US Air (PHL-US)	9.13158	2.85095

Table 2. (Continued) Estimates of the Rosse-Panzar Test Statistic for Inbound Traffic, Ranked from Lowest to Highest

City - Pair	RP-Statistic (inbound)	Standard errors*
24 Minneapolis St Paul/Wold-Chamb.- Northwest (MSP-NW)	9.17014	1.70015
25 Newark Int. - Continental (EWR-CO)	10.2999	3.93423
26 Dallas Ft. Worth - American (DFW-AA)	13.2785	1.85012
27 G. Bush Intl. Continental/Houston - Continental (IAH-CO)	14.8425	2.17619
28 Chicago O'Hare Intl. - United (ORD-UA)	16.6315	4.24272
29 Dallas Ft. Worth - Delta (DFW-DL)	18.6381	3.86058

Note: *All coefficients have a significantly positive test statistic, which is also significantly different from one.

Tables 1 and 2 present the Rosse-Panzar test statistic and its standard error for the 29 airport-pairs by outbound traffic and inbound traffic, respectively. In our empirical testing for Rosse-Panzar and for cross-sectional regressions in the next section, we employ quarterly price indices constructed from raw data provided by the DOT's Form 41 as Air Carrier Financial Statistics, and Air Carrier Traffic Statistics. The price indices for labor, fuel, and materials are constructed using index number theory. The price of capital in contrast is constructed by employing the Capital Asset Pricing Model (CAPM). The CAPM computes the correct risk-adjusted return for a risky asset within the framework of mean-variance portfolio theory. Since it provides an economic measure of the price of capital and reflects the true risk-adjusted opportunity cost, it is vastly superior to conventional accounting measures for the price of capital.[4] Price data were derived from Database 1A of the DOT's origin and destination survey (O&D). The sample period

[4] For a more detailed discussion of how the price of capital is calculated, see Fischer and Kamerschen (2002).

covers the 24 quarters between the first quarter of 1991 and the fourth quarter of 1996.

Church and Ware (1999) point out that the Rosse-Panzar test shows what the market structure or degree of monopoly is not and does not suggest what is. Following this approach, we can rule out monopoly and perfect competition for all airport-pairs that have a significantly positive test statistic, which is also significantly different from 1. This is clearly the case for the majority of the airport-pairs. Thus, the finding for these airport-pairs is consistent with the structural model, which indicates conduct somewhere in between the collusive solution, i.e. monopoly, and perfect competition. A closer look at the airport-pairs with significantly negative estimates for the test statistic is warranted. Recall that a negative test statistic can imply both competition or monopoly. The airport-pairs that require closer scrutiny are Delta in the Detroit market (inbound only), Memphis, Miami, Chicago O'Hare, and Philadelphia; United in the Washington-Dulles market, US Air in Pittsburgh (inbound only) and American for Miami and Chicago O' Hare. Any further investigation into market structure with the Rosse-Panzar test statistic remains inconclusive. Finally, the magnitude of the estimates seems too large if one wants to follow Shaffer's suggestion regarding the estimation of the Lerner index. The estimates obtained seem to preclude this estimation. However, the estimates are very robust to changes in the specification of the model. Any potential explanation of the magnitude of the estimates will have to explore in greater detail two assumptions that could lead to implausibly high values for the test statistic. The first is the assumption that the air carrier is a price taker on the input side. There is some evidence that this is not the case, particularly for the input labor. Heavy unionization and widespread collective bargaining suggest that airlines face a less than competitive market for their labor inputs. The second is the assumption that the industry is in long-run equilibrium. Recall that such an assumption is crucial for the Rosse-Panzar test to work. Shaffer (1982a, b) explicitly points to the almost contradictory nature of the assumptions that all observations are identified, and controlled for as being in long-run equilibrium. In particular, when working with a time-series sample like the airport-pair markets, any change in factor prices involves some adjustment, which is unlikely to be completed exactly by the end of the observed period. However, it is precisely this variation in prices that is needed to identify the test statistic.

B. A Cross-Section Regression

This section presents a different approach to the investigation of pricing strategies employed by airlines. The section develops a cross-section regression model employing price data and route characteristics for a cross-section sample of airline routes originating in Atlanta. The objective is to assess how particular route characteristics affect the price on a given route. In developing the model, we closely follow Peteraf and Reed (1994) and Borenstein (1989), adjusting the model according to the requirements of the investigation and availability of data. Observations are for the four quarters of 1996. Each observation consists of one carrier serving one airport-pair. Both nonstop and one-stop service are included. The equation to be estimated is specified as follows

$$\begin{aligned}\ln \mathit{YIELD} = \; & a_0 + b_1 \ln \mathit{PASSENGER} + b_2 \ln \mathit{DISTANCE} \\ & + b_3 \ln \mathit{AVERANGE\ COST} + b_4 \ln \mathit{INCOME} \\ & + b_5\, \mathit{MARKETSHARE} + b_6\, \mathit{HHI} + b_7\, \mathit{VALUJET} \\ & + b_8\, \mathit{VACATION}\end{aligned} \tag{9}$$

where YIELD is defined as price divided by distance. That is, YIELD measures the average fare charged by the observed carrier on the given route, divided by stage length so as to obtain the price per mile and normalize across different stage lengths. PASSENGERS is equal to the number of passengers transported on the route during a quarter. It measures the total number of all local origin-to-destination passenger. DISTANCE measures the stage length between the departure and arrival cities. AVERAGECOST is a proxy for the cost-competitiveness of the airline offering the service and is measured in average cost per seat mile. Adjustments are made to account for different average stage lengths across carriers. INCOME is a measure of disposable personal income for the metropolitan statistical area of the destination. It is included to capture aggregate income at the destination. MARKETSHARE captures the market share that the airline commands on a given route. It measures the share of all local origin-to-destination passengers for the observed carrier on a given route. Thus, it is constructed by dividing PASSENGERS by the total number of local origin-to-destination passengers. HHI is the Herfindahl-Hirschman index for the route under consideration; it ranges from 0 to 1. Finally VALUJET is an indicator variable taking the value of one if a particular airport-pair is served by ValuJet airlines and zero otherwise. It is designed to measure whether the presence of a discount carrier has a depressing effect on prices. Finally, VACATION is a dummy variable indicating whether a destination is primarily a vacation spot. Price data are obtained from the DOT's origin and destination (O&D) survey for the four quarters of 1996, along with information on passengers. The O&D survey also indicates whether ValuJet is serving a particular airport-pair market. Using the quantity data, the measures for market share and concentration are constructed. Distance is taken from Delta Air Line's worldwide timetable, effective June 1, 1997. Data on population and income for the Metropolitan Statistical Areas have been compiled by the Bureau of Labor statistics.

The expected sign for PASSENGERS is negative since with a larger number of passengers the load factor increases, and therefore unit costs per passenger should decrease. DISTANCE is one of the most important determinants of airline cost. As distance increases, cost per mile decreases as discussed previously. Since aircraft burn most fuel during take-off and landing, and fixed cost can be spread over more miles, we expect unit cost per mile to decrease as stage length increases. Therefore, the overall effect of DISTANCE on YIELD is hypothesized to be negative. AVERAGECOST serves as a proxy for a carrier's cost efficiency. AVERAGECOST is calculated for the entire domestic system, but adjusted with respect to distance. For example a carrier with relatively high system-wide average cost, but a short average stage length may still be more cost efficient than a carrier with slightly lower average cost, but longer average stage length. The adjustment renders the AVERAGECOST proxies comparable for any given route. The expected sign for AVERAGECOST is positive, since less efficient firms are hypothesized to demand higher fares. Since air travel is a normal good, an increase in disposable income should increase the price of air travel. Hence, the sign for INCOME is expected to be positive. Controlling for concentration, a firm with a higher market share is expected to realize a higher yield. Therefore, the expected sign for MARKETSHARE is positive. The sign for HHI is theoretically ambiguous. A dominant firm could find it more convenient and easier to maintain high prices if it competes against a fringe of small firms rather than a fairly large and well-established rival. In the first scenario the HHI would be smaller than in the second. The predicted sign would be negative. However, holding market share constant, a higher HHI may make it more feasible for firms to collude, hence raising prices. On the other hand if dominance stems from technological advantages of the dominant firm such as cost efficiency or effective marketing, rather than anti-competitive conduct, yields for other firms should decrease. In the former case the sign is positive, whereas the latter scenario suggests a negative sign. Overall, the sign depends on the sources of concentration. The presence of a lowcost competitor such as ValuJet in any given market should provide for increased and more vigorous competition, and therefore should bring yields down. Therefore, the expected sign for VALUJET is negative. Finally, leisure travelers are more price sensitive; their demand for air travel is consequently more elastic. A market to a destination that comprises a large share of leisure travelers therefore should, ceteris paribus, afford lower yields. The portion of leisure travelers is assumed to be higher on routes to vacation spots. Therefore, the hypothesized sign for VACATION is negative.

Before we carried out the regression, some econometric issues were addressed. First there is a potential problem regarding the possible endogeneity of PASSENGERS, MARKETSHARE, and HHI. Indeed, a Haussmann specification test rejects exogeneity for PASSENGERS and MARKETSHARE. Therefore, we proceed with estimation using instruments and 2-stage least squares. As the preferred set of instrument, we include all the exogenous variables and their interactions with the dummies, as well as the carrier's share of all origin and destination passengers in Atlanta. We also include the overall population of the destination's metropolitan area, its square, and distance squared.

Table 3 presents the coefficient estimates, along with their standard errors. All coefficients have the expected sign where there existed unambiguous predictions regarding the sign. Moreover, all coefficient estimates are highly significant at better than the one-percent level. The coefficient estimates imply that a 10 percent increase in local origin-and destination passengers decreases fares by 1 percent. An increase in distance by 10 percent decreases fares by 7

Table 3. Cross-Section Regression Parameter Estimates for the Dependent Variable Yield

Variable	Coefficient	Standard error*
CONSTANT	3.63783	0.111927
PASSENGERS	-0.099404	0.00711
DISTANCE	-0.702309	0.015092
MARKETSHARE	1.00035	0.058172
HERFINDAHL-HIRSCHMAN INDEX (HHI)	-0.347235	0.042411
AVERAGE COST	0.332542	0.055287
INCOME	0.15879	0.034166
VACATION	-0.121219	0.017988
VALUJET	-0.160558	0.015952
R^2	0.774	

Note: * Examining the p-values corresponding to the appropriate t-value shows that all coefficients are significant at the 1% or better level.

percent on average. Furthermore, a one-point increase in the observed carrier's market share increases fares by 1 percent. Moreover, the estimates suggest that a 10 percent increase in average cost translates into a 3.3 percent increase in fare. The income elasticity of demand is approximately 16 percent. An increase in concentration as measured by the HHI index reduces the yield. Therefore, the model suggests that the dominant carrier Delta enjoys technological advantages over its rivals or that there is some degree of competition provided by another carrier. Most important for advocates of vigorous competition is the coefficient for VALUJET, indicating that fares in airport-pair markets served by ValuJet were on average 16 percent lower than on routes where such competition was absent. This is a ringing endorsement for low-cost carriers. It strongly suggests that in the interest of the traveling public, competition in the airline industry should be encouraged, promoted, and facilitated wherever possible.

IV. Conclusions

We employ a reduced form model called the Rosse-Panzar test to calculate price-cost margins in selected airport-pair markets originating from Atlanta. The statistics are generally positive and quite large, indicating that carriers are neither in perfect competition nor perfectly colluding. Unlike structural models, the Rosse-Panzar test is only sufficiently powerful to reject certain outcomes of market conduct. We find that in all airport-pairs, the existence of the Bertrand outcome, which is equivalent to perfect competition, is resoundingly and consistently rejected, as is the outcome describing perfect collusion, which is equivalent to the joint monopoly outcome.

In contrast, the Cournot solution cannot be rejected. In most markets, conduct is consistent with the Cournot solution. However, the Rosse-Panzar test is not powerful enough to identify a specific model of conduct. Our findings show that conduct in most airport-pairs is also consistent with a range of conduct deviating from the Cournot oligopoly both to the more and less competitive behavior. That is, conduct is consistent with a wide range of intermediate solutions between the monopoly outcome and perfect competition. A cross-section pricing regression model to study pricing behavior supplements the Rosse-Panzar approach. We find that all variables affect the dependent variable as hypothesized and that all parameter estimates are highly significant. We find that yield or price per mile traveled is positively correlated with the airline's average costs, its market share in a given airport-pair market and the income in the metropolitan area where the airport is located. Yield is negatively correlated with enplaned passengers, since, as the load factor rises, the cost per passenger is declining. It is negatively correlated with the Herfindah-Hirschmann-Index for a given market and with the distance between airports. It is also significantly lower in markets that are considered primarily destinations for vacationers. Most importantly, we find that the presence of lowcost competition has a significant and substantial impact on average yields. For 1996, the period under investigation, other things being the same, average fares were about 16 percent lower in markets where ValuJet was present than in those in which it did not operate. In summary, we find sufficient evidence that the industry, at least as it relates to airport-pair markets originating from Atlanta, has some way to go to reach the benchmark of perfect competition.

References

Air Carrier Financial Statistics, reported on Form 41, provided by Database Products, Dallas, TX., 1991:1-1996:4.

Ashenfelter, O., and D. Sullivan (1987), "Nonparametric Tests of Market Structure: An Application to the Cigarette Industry," *Journal of Industrial Economics* **35**: 483-498.

Bailey, E.E. (2002), "Aviation Policy: Past and Present," *Southern Economic Journal* **69**: 12-20.

Berry, S. (1992), "Estimation of a Model of Entry in the Airline Industry," *Econometrica* **60**: 889-917.

Borenstein, S. (1989), "Hub Dominance and High Fares: Airport Dominance and Market Power in the U.S. Airline Industry," *Rand Journal of Economics* **20**: 344-365.

Church, J., and R. Ware (1999), *Industrial Organization: A Strategic Approach*, Boston, Massachusetts, Irwin McGraw-Hill.

Fischer, T., and D.R. Kamerschen (2002), "Price-Cost Margins in the Airline Industry Using a Conjectural Variation Approach," *Journal of Transport Economics and Policy,* forthcoming.

Gowrisankaran, G. (2002), "Competition and Regulation in the Airline Industry," *Federal Reserve Bank of San Francisco Economic Letter 2002-01*: 1-3.

Nathan, A., and E.H. Neave (1989), "Competition and Contestability in Canada's Financial System: Empirical Results," *Canadian Journal of Economics* **22** (3): 576-594.

Panzar, J.C., and J.N. Rosse J.N. (1987), "Testing for 'Monopoly' Equilibrium," *Journal of Industrial Economics* **35**: 443-457.

Peteraf, M., and R. Reed (1994), "Pricing and Performance in Monopoly Airline Markets," *Journal of Law and Economics* **37**: 193-213.

Rosse, J.N., and J.C. Panzar (1977), "Chamberlin vs. Robinson: An Empirical Test for Monopoly Rents," *Bell Laboratory Economics Discussion Paper 90*, and *Stanford University Studies in Industry Economics* 77.

Seade, J.E. (1980), "On the Effects of Entry," *Econometrica* **48**: 479-489.

Shaffer, S. (1982a), "Competition, Conduct and Demand Elasticity," *Economics Letters* **10**:167-171.

Shaffer, S. (1982b), "A Nonstructural Test for Competition in Financial Markets," Proceedings of a Conference on Bank Structure and Competition, Federal Reserve Bank of Chicago, April 12-14: 225-243.

Shaffer, S. (1983a), "The Rosse-Panzar Statistic and the Lerner Index in the Short Run," *Economic Letters* **11**: 175-178.

Shaffer, S. (1983b), "Non-Structural Measures of Competition Toward a Synthesis of Alternatives," *Economics Letters* **12**: 349-353.

Shaffer, S., and J. Di Salvo (1994), "Conduct in a Banking Duopoly," *Journal of Banking & Finance* **18**: 1063-1082.

Sharpe, W. F. (1964), "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *Journal of Finance* **19**: 425-442.

Silberberg, E. (1974), "The Theory of the Firm in 'Long-Run' Equilibrium," *American Economic Review* **84**: 734-741.

Sullivan, D. (1985), "Testing Hypotheses About Firm Behavior in the Cigarette Industry," *Journal of Political Economy* **93**: 586-598.

United States Department of Transportation (DOT), "Origin and Destination Survey," Database 1A. Hub-Study ATL, provided by Database Products, Inc., Dallas, TX., 1991:1-1996:4.

Journal of Applied Economics, Vol. VI, No. 1 (May 2003), 95-125

TRADE LIBERALISATION WITH COSTLY ADJUSTMENT

ALVARO FORTEZA AND ROSSANA PATRÓN[*]
Universidad de la República, Uruguay

Submitted July 2001; accepted June 2002

The paper analyses the efficiency and the distributional effects of eliminating a tariff in a protected sector, in a Heckscher-Ohlin model of trade with costs of adjustment. The tariff can be eliminated at the onset or after a while. In case of postponing it the government may pre-announce the policy change or may not do it and surprise the private sector. It is shown that while large adjustment costs reduce the efficiency gains from trade liberalisation, small to moderate adjustment costs may *raise* the efficiency gains from a pre-announced liberalisation. The adjustment costs reduce the effects on factor returns from a sudden unanticipated liberalisation. The distributional effects of trade liberalisations are more complex when the policy is pre-announced. For small and moderate levels, the adjustment costs may increase the effects of the policy on factor returns. Also, the "value of the announcement" rises with the adjustment costs.

JEL classification codes: F110, F130
Key words: adjustment costs, trade liberalisation

I. Introduction

This paper analyses the welfare gains and losses from the elimination of tariffs in the presence of costs of adjustment, using a dynamic extension of an otherwise standard Heckscher-Ohlin (HO) model of trade. The paper compares different alternatives of trade liberalisation, including a sudden unanticipated elimination of the tariffs, a pre-announced elimination of the tariffs, and a

[*] Departamento de Economía, Facultad de Ciencias Sociales, Universidad de la República, Uruguay. José E. Rodó 1854, CP 11200, Montevideo, Uruguay. Email: Alvarof@decon.edu.uy, Rossana@decon.edu.uy. We thank an anonymous referee from this journal for useful comments. The usual disclaimer applies.

postponed, but still not announced, elimination of the tariffs. We analyse both the efficiency and the distributional effects of the trade policy. The efficiency effects are measured as the response in the welfare of the representative agent in a homogeneous-society version of the model, and the distributional effects are measured by the welfare gains and losses of different individuals in a heterogeneous-society version of the model.

Costs of adjustment arise from many sources, including hiring, firing, and training labour, installing and adapting machines and buildings, and doing marketing and adapting the production distribution nets. With these so many sources of costs of adjustment, it is not obvious how the adjustment costs function should be specified. Furthermore, there is now an extensive literature showing that the economic dynamics associated to costly adjustment does depend on some details of the specification of the adjustment costs function. In one vein, some authors have emphasised the relevance of distinguishing net from gross adjustment costs (Hamermesh, 1993; Hamermesh et al., 1994). The former arises when the level of employment is changed, and the latter occur whenever workers are hired or fired, even if the level of employment remains unchanged. A similar distinction has been made for capital (Neary, 1978; Grossman, 1983; Clarete, et al., 1994). Gross adjustment costs give rise to sector specificity and to different returns of the same production factor across sectors.

In a related but different vein, the literature has explored the effects of adding fixed adjustment costs, non convex adjustment costs, and marginal adjustment costs that do not tend to zero as the input change tends to zero (Oi, 1962; Rothschild, 1971; Kemp and Wan, 1974; Hamermesh, 1989; among others). This literature has shown that these adjustment cost functions may give place to very different responses to price shocks, ranging from no response at all to minor shocks, to immediate one-period adjustment.

We adopt a quadratic adjustment cost function, in the fashion of Sargent (1978). In so doing, we make several choices. First, we focus on net adjustment costs, leaving aside the costs stemming from turnover. Factors can be costlessly moved from one sector to the other, and hence the return to production factors is equalised across sectors. In this respect, we keep close to the standard HO model. But because of the cost of changing the level of production, competitive firms make non-zero profits. Hence, unlike previous models of trade liberalisation, the model in this paper exhibits changes in the value of the firms associated to trade reforms. Besides, these changes are different across sectors. In the real world, structural changes in which some sectors expand and some other sectors contract seem to be associated to significant changes in the values of the involved firms. Our model may be useful to analyse this aspect of the liberalisation process that has received little attention in the literature. Second, quadratic adjustment costs leave out of our analysis issues of hysteresis and lumpy responses to shocks. Admittedly, these issues are likely to be important in the real world. We leave them aside because we want to preserve the HO characteristics of the model in the steady state, while having a gradual adjustment process during the transition.

More often than not trade reforms come as a building block of a broader package of structural reforms that include deregulations, macroeconomic stabilisation, financial liberalisation, capital account liberalisation, and privatisation. The question then arises about the optimal sequencing of the reforms in these different areas. The extensive literature that deals with this issue has come with no simple policy recipe.[1] We make no attempt to provide a general answer to this largely unsettled issue; the model in this paper is too simple to deal with most of the effects that must be taken into account in any comprehensive assessment of the sequencing of reforms. Notwithstanding, our model does have some implications for the sequencing of trade liberalisation and deregulations affecting adjustment costs. We show that, in the case of pre-announced liberalisations, it could be optimal to postpone deregulations that reduce (moderate) adjustment costs until tariffs have been eliminated.

Adjustment costs have played an important role in informal arguments that have been put forward to support the gradualist view on trade liberalisation (see for instance, Michaely, 1986). Our analysis shows that net adjustment costs provide no reason for delay, and hence the gradualist view must be based on rigidities that cannot be appropriately represented with this type of adjustment costs. We briefly review some of these sources of rigidity in the next paragraph.

[1] See, among others, Choksi and Papageorgiou (1986), Edwards (1989), Funke (1994), Edwards (1994).

Being our goal analytic, we decided to focus on a narrow set of issues, keeping the model as close as possible to the HO tradition, hence leaving aside many important considerations that should not be dismissed in a balanced assessment of trade reform. Concerns about unemployment are usually prominent in the policy debate about trade liberalisation, despite of some recent empirical literature indicating that the short run effects of trade liberalisation on unemployment may be small (Papageorgiou et al., 1991; Edwards, 1994). Early analytical treatments of this issue can be found in Neary (1982) and Mussa (1986). Several episodes of trade liberalisation were associated to large current account deficits and consumption booms. These distortions have been explained in terms of the lack of credibility of the liberalisation process, or the hypothesis that agents think that the tariff reduction may be temporary (Calvo, 1988; Calvo and Mendoza, 1994). Karp and Paul (1994) analyse the optimal timing of trade reform in the presence of congestion costs. They argue that because of congestion externalities, private and social marginal adjustment costs may differ, and reallocation tends to occur too rapidly. Nevertheless, they show that trade reforms should begin with trade liberalisation, and only if the government has commitment capacity there should be an intermediate phase with positive tariffs, followed by full liberalisation. Investment decisions are usually costly to reverse. Coupled with uncertainty, irreversibility may give rise to substantial inertia and hysteresis (for a survey, see Dixit, 1992). Albuquerque and Rebelo (1998) explore the implications of irreversible investment and uncertain duration of the trade reform for the performance of the economy in the aftermath of the trade liberalisation reform.

The paper proceeds as follows. In Section II, we present and solve the formal model. In Section III, we report the main results from simulations. Section IV concludes with some final remarks.

II. The Model

A. Production and Income

There are two productive sectors that use two factors of production, capital and labour. The technology is assumed Cobb Douglas:

$$F_i\,(L_{i,t}, K_{i,t}) = H_i\,(K_{i,t})^{\alpha_i}\,(L_{i,t})^{1-\alpha_i} \qquad i = A, B \tag{1}$$

Competitive firms rent capital paying return r_t per unit of capital to owners of capital. Firms also hire labour, paying a wage w_t to workers, and incurring in quadratic adjustment costs when the total amount of labour occupied in the firm is changed. With only net adjustment costs, there is no significant difference between labour and capital adjustment costs. Indeed, we are assuming that there is a cost associated to changing the level of production. For ease of computation, we write it as a cost of changing the employment of a production factor, but it can be shown that there is an equivalent formulation in terms of the other production factor and still another equivalent formulation in terms of output.

Individual firms do not control prices of production factor services nor prices of goods $P_{i,t}$. Entrepreneurs in sector i choose the path of labour and capital to maximise the value of the firm:[2]

$$\underset{\{L_{i,t}, K_{i,t}\}\ t=0,\ldots,\infty}{Maximise} \quad \sum_{t=0}^{\infty}\left[\left(P_{i,t}Y_{i,t} - w_t L_{i,t} - r_t K_{i,t}\right)\Big/\prod_{s=0}^{t}\left(1+R_s\right)\right] \tag{2}$$

$s.t.$

$$Y_{i,t} = F_i\,(L_{i,t}, K_{i,t}) \ - \frac{a_i}{2}\left(L_{i,t} - L_{i,t-1}\right)^2$$

$$0 \le L_{i,t} \le \overline{L}; \quad 0 \le K_{i,t} \le \overline{K}$$

$$L_{i,-1} \quad given$$

where a_i is the adjustment cost parameter in sector i,$\overline{L}$ and $\overline{K}$ are the factor endowments, $L_{i,-1}$ is the initial allocation of labour, and R_s is the interest rate. The first order conditions are:

[2] In order to simplify notation the same symbols represent both the employment of the firm and that of the whole sector.

$$P_{i,t} H_i (1-\alpha_i)\left(\frac{K_{i,t}}{L_{i,t}}\right)^{\alpha_i} = w_t + a_i P_{i,t}\left(L_{i,t} - L_{i,t-1}\right) - \frac{a_i P_{i,t+1}}{1+R_{t+1}}\left(L_{i,t+1} - L_{i,t}\right); \qquad t = 0,\ldots\infty \tag{3}$$

$$P_{i,t} H_i \alpha_i \left(\frac{K_{i,t}}{L_{i,t}}\right)^{\alpha_i - 1} = r_t; \qquad t = 0,\ldots\infty \tag{4}$$

In the tradition of the Heckscher-Ohlin model, we assume that factor endowments in the economy are fixed. There is no capital accumulation, and no demographic growth. Markets are competitive and prices are fully flexible, so the markets for production factors clear in every moment:

$$L_{A,t} + L_{B,t} = \overline{L}; \qquad t = 0,\ldots\infty \tag{5}$$

$$K_{A,t} + K_{B,t} = \overline{K}; \qquad t = 0,\ldots\infty \tag{6}$$

The economy is small. Domestic events do not modify international prices $P^*_{i,t}$, but the government sets taxes and subsidies on foreign trade $\tau_{i,t}$ that alter domestic prices (the foreign exchange rate is normalised to 1):

$$P_{i,t} = P^*_{i,t}\left(1 + \tau_{i,t}\right) \tag{7}$$

There is no international borrowing and lending. The interest rates are determined to clear domestic credit markets (see next section).

Equations (3) to (7) define a system of non-linear second-order difference equations, that can be solved for eight endogenous variables: $L_{A,\,t,}$ $L_{B,\,t}$, $K_{A,\,t,}$ $K_{B,\,t}$, r_t, w_t, $P_{A,\,t}$ and $P_{B,\,t}$. Two points in the path of each of the two dynamic variables ($L_{A,\,t}$ and $L_{B,\,t}$) must be given to pin down a particular solution. It is natural to set the initial level of employment, $L_{A,\,-1}$ and, $L_{B,\,-1}$ as one of those points. Infinite paths are still consistent with both the system (3) to (7) and initial employment, but the saddle path dynamics of this system imply that firms can rule out all save one path, the one converging to the steady state. Other paths are diverging and eventually violate the employment constraints in the firms' programs $\left(0 \le L_{i,\,t} \le \overline{L}\right)$. Rationality hence implies that the economy eventually converges to the steady state. Output in both sectors can be computed using the paths of capital and labour and equation (1).

Profits are zero in the long run, but not during the transition. In the steady state, when employment stabilises, production factors are paid their marginal product (see equations (3) and (4)). This result and the assumption of constant returns to scale imply zero profits in the long run. During the transition, adjustment costs operate as barriers to entry and exit and firms make profits or loses. Accordingly, there is a value attached to the firm. Interestingly, the simulation results presented below show that there is no simple relationship between the performance of the sector, as measured by output or employment, and the value of the firms. Depending on the timing of the announcements and the implementation of trade liberalisation, firms in the contracting sectors may make loses or profits.

B. Consumption, Interest Rates and Foreign Trade

We develop two versions of the model, one with homogeneous and the other with heterogeneous population. The representative agent version of the model allows us to focus on the efficiency effects of trade liberalisation, postponing the analysis of the distributional effects of this policy. The heterogeneous population version of the model assumes that the property rights over the production factors and the firms are non-uniformly distributed in the population. The productive sector is the same in both versions. Like in the static HO model, the productive decisions do not depend on the distribution of the property rights over production factors. We present the representative agent version first and the heterogeneous population model later in this same section.

B.1. The Representative Agent Model

The economy is populated by a constant number of identical and infinitely lived individuals. In order to simplify notation, the size of the population is normalised to 1. The same symbol represents both the aggregate and the individual variables. Individuals own the production factors and the firms. Hence, both the returns of the production factors and the benefits of the firms add to individuals' income, and this sum equals gross revenues of the firms $\left(r_t \overline{K} + w_t \overline{L} + Benefits = P_{A,t}Y_{A,t} + P_{B,t}Y_{B,t}\right)$. Individuals also receive a uniform lump-sum transfer from the government b_t.[3] To keep as close as possible to the conventional HO model, we get rid off accumulation of goods by assuming that both goods are perishable. Individuals can accumulate net financial assets A_t, borrowing and lending at the interest rate R_t.

The utility function is additively separable in time, with discount factor β. Per period utility is Cobb-Douglas in consumption of both goods.

$$\underset{\{C_{A,t},C_{B,t}\}\ t=0,\dots,\infty}{Maximise} \quad \sum_{t=0}^{\infty} \beta^t \, C_{A,t}^{\theta} \, C_{B,t}^{1-\theta} \tag{8}$$

$s.t.$

$$P_{A,t}C_{A,t} + P_{B,t}C_{A,t} + A_{t+1} \leq P_{A,t}Y_{A,t} + P_{B,t}Y_{B,t} + b_t + A_t\left(1+R_t\right);\ t \qquad t = 0,\dots,\infty$$

This program yields corner solutions, in terms of the choice of present versus future consumption, for most combinations of values of parameters and of exogenous variables. These solutions imply that the consumer chooses either to consume all his wealth in the first period and nothing therein or, in the other extreme, to indefinitely postpone consumption. In the first case, all families would want to borrow in the first period and the credit market would be in excess demand. The interest rate would necessarily rise. In the second extreme case, all families would want to lend so there would be an excess supply of loans. The interest rate would fall. There is an intermediate value of the interest rate such that individuals' plans can be consistent in the aggregate. We derive the expression for this equilibrium interest rate in the appendix, and reproduce it here as:

[3] This assumption is discussed in the following section.

$$1 + R_t = \left(\frac{P_{A,t}{}^{\theta} \; P_{B,t}{}^{1-\theta}}{P_{A,t-1}{}^{\theta} \; P_{B,t-1}{}^{1-\theta}} \right) \frac{1}{\beta} \tag{9}$$

$$= \left(1 + \frac{P_t - P_{t-1}}{P_{t-1}} \right) (1 + \mathit{subjective\, discount\, rate})$$

Therefore, the equilibrium real interest rate equalises the subjective discount rate, with the real interest rate computed with the relevant price index for this economy $\left(P_t = P_{A,t}{}^{\theta} P_{B,t}{}^{1-\theta} \right)$.[4]

Two different consumption decisions are embedded in program (8). One is an intratemporal decision: how much to consume of each good within each period. The first order conditions indicate that the composition of the consumption basket in each period must be determined according to the following rule:

$$\frac{C_{B,t}}{C_{A,t}} = \left(\frac{1-\theta}{\theta} \right) \frac{P_{A,t}}{P_{B,t}}; \quad t = 0,\ldots,\infty \tag{10}$$

The other decision consumers must make is intertemporal in nature: how much to consume today and how much tomorrow. Consumers are indifferent between consuming today or tomorrow, when the interest rate satisfies equation (9) (see the Appendix for the details). Hence, individual consumption is not fully determined by program (8).

Goods markets are in equilibrium when output plus net imports $M_{i,\,t}$ equal domestic consumption. There is no accumulation of goods, for goods are assumed perishable.

$$Y_{i,t} + M_{i,t} = C_{i,t}; \quad t = 0,\ldots,\infty \tag{11}$$

The assumption that there is no international credit implies that the current account of the balance of payments must be balanced:

[4] This solution depends on the particular assumptions about the utility function (see, for instance, Sargent, 1988).

$$P^{*}_{A,t} M_{A,t} + P^{*}_{B,t} M_{B,t} = 0; \quad t = 0,...,\infty \tag{12}$$

The system of equations (10) to (12) determine consumption and net imports in both sectors, given prices and output.

B.2. The Heterogeneous Population Model

Individuals in this economy may receive income from five different sources: wages, returns to capital, profits of firms in sector A, profits of firms in sector B, and transfers from the government. Individual ‘h’ solves the following program:

$$\underset{\left\{C^h_{A,t}, C^h_{B,t}\right\}\ t=0,...,\infty}{\mathit{Maximise}} \quad \sum_{t=0}^{\infty} \beta^t \, C^h_{A,t} \, {C^h_{B,t}}^{1-\theta} \tag{13}$$

$s.\ t.$

$$P_{A,t} C^h_{A,t} + P_{B,t} C^h_{B,t} + A^h_{t+1} \leq r_t K^h_t + w_t L^h_t + B^h_{A,t} + B^h_{B,t} + b^h_t + A^h_{t+1}(1 + R_t);$$

$$t = 0,...,\infty$$

where $B^h_{i,t}$ are the profits that agent ‘h’ makes from the property of firms in sector i. Adding the individual budget constraints over ‘h’ gives the representative agent resource constraint in equation (8).

Equations (9) and (10) continue to hold, and hence the consumption basket has the same composition for all consumers. The difference is in the level: consumers with more resources will enjoy larger consumption. We use these properties in the simulations below to compute the welfare gains from different groups of individuals.

C. The Government

The government sets taxes and subsidies on foreign trade, driving a wedge between domestic and foreign prices. The proceeds of net taxes on foreign trade are distributed uniformly among individuals in a lump-sum fashion.

Hence, the government budget is balanced in each period. This assumption allows us to focus on the straight effects from trade policy.

$$\tau_{A,t} P^{*}_{A,t} M_{A,t} + \tau_{B,t} P^{*}_{B,t} M_{B,t} = b_t \tag{14}$$

Note that $\tau_{i,t}$ represent several trade policy instruments. It is an *import tariff* if $M_{i,t} > 0$ and $\tau_{i,t} > 0$; it is an *import subsidy* if $M_{i,t} > 0$ and $\tau_{i,t} < 0$; it is an *export tax* if $M_{i,t} < 0$ and $\tau_{i,t} < 0$; and it is an *export subsidy* if $M_{i,t} < 0$ and $\tau_{i,t} > 0$. Taxes and subsidies on foreign trade are policy instruments, while the lump-sum transfers are endogenously determined by the government budget (14).

D. The Phase Diagram

The qualitative properties of the model can be analysed with the help of a phase diagram. The model exhibits saddle path dynamics, and the steady state is the standard static HO equilibrium. Equations (4) to (6) imply that:

$$\begin{aligned} P_{A,t} H_A \alpha_A \left(\frac{K_{A,t}}{L_{A,t}} \right)^{\alpha_A - 1} &= P_{A,t} H_A \alpha_A \left(\frac{\overline{K} - K_{B,t}}{\overline{L} - L_{B,t}} \right)^{\alpha_A - 1} \\ &= P_{B,t} H_B \alpha_B \left(\frac{K_{B,t}}{L_{B,t}} \right)^{\alpha_B - 1} \\ &= P_{B,t} H_B \alpha_B \left(\frac{\overline{K} - K_{A,t}}{\overline{L} - L_{A,t}} \right)^{\alpha_B - 1} \end{aligned} \tag{15}$$

These equations define two implicit functions mapping employment into capital in each sector:

$$K_{i,t} = K_i \left(L_{i,t} \right);\ i = A, B \tag{16}$$

with first derivaties:

$$\frac{dK_{A,t}}{dL_{A,t}} = \frac{dK_{B,t}}{dL_{B,t}} = \left(\frac{(1-\alpha_A) L_{B,t} + (1-\alpha_B) L_{A,t}}{(1-\alpha_A) K_{B,t} + (1-\alpha_B) K_{A,t}} \right) \frac{K_{A,t} K_{B,t}}{L_{A,t} L_{B,t}} > 0 \tag{17}$$

The fundamental dynamic equation of the model follows from equations (3), (5) and (16):

$$P_{A,t}\,H_A\left(1-\alpha_A\right)\left(\frac{K_A\left(L_{A,t}\right)}{L_{A,t}}\right)^{\alpha_A} - P_{B,t}\,H_B\left(1-\alpha_B\right)\left(\frac{\overline{K}-K_A\left(L_{A,t}\right)}{\overline{L}-L_{A,t}}\right)^{\alpha_B} \qquad (18)$$

$$=\left(a_A P_{A,t} + a_B P_{B,t}\right)\left(L_{A,t} - L_{A,t-1}\right) - \frac{\left(a_A P_{A,t+1} + a_B P_{B,t+1}\right)}{1+R_{t+1}}\left(L_{A,t+1} - L_{A,t}\right)$$

This non-linear-second-order difference equation in employment determines a family of integral curves. Two additional conditions are needed to pin down a particular solution to equation (18). One is the initial level of employment. The other is a transversality condition, implicit in the feasibility constraint that employment in any sector is non negative and smaller than or equal to total labour supply. It is shown below that all save one path eventually violate this feasibility constraint.

It proves useful to write equation (18) as a first-order system in the level and the first difference of employment:

$$P_{A,t}\,H_A\left(1-\alpha_A\right)\left(\frac{K_A\left(L_{A,t}\right)}{L_{A,t}}\right)^{\alpha_A} - P_{B,t}\,H_B\left(1-\alpha_B\right)\left(\frac{\overline{K}-K_A\left(L_{A,t}\right)}{\overline{L}-L_{A,t}}\right)^{\alpha_B} \qquad (19)$$

$$=\left(a_A P_{A,t} + a_B P_{B,t}\right)\left(L_{A,t} - L_{A,t-1}\right) - \frac{a_A P_{A,t+1} + a_B P_{B,t+1}}{1+R_{t+1}}\,X_t$$

$$X_{t-1} = L_{A,t} - L_{A,t-1} \qquad (20)$$

The phase diagram of this system will be represented in $(L_{A,\,t-1}, X_{t-1})$. We will first derive the phase line for constant employment (and the consequence dynamics) and then the phase line for constant variation of employment (and its respective dynamics).

(i) The locus of constant employment, $L_{A,\,t} = L_{A,\,t-1}$. Equation (20) imply that this locus is $X_{t-1} = 0$.

(ii) Dynamics of employment,

$$\Delta L_A = L_{A,t} - L_{A,t-1} > (=,<)\ 0 \quad if \quad X_{t-1} > (=,<)\ 0$$

(iii) The locus of constant variation of employment, $X_t = X_{t-1.}$ The condition that defines this locus is: $X_t = X_{t-1} = L_{A,t} = L_{A}{,}_{t-1}$; using this condition in (19):

$$P_{A,t}\, H_A\left(1-\alpha_A\right)\left(\frac{K_A\left(L_{A,t-1}+X_{t-1}\right)}{L_{A,t-1}+X_{t-1}}\right)^{\alpha_A}$$

$$-\,P_{B,t}\, H_B\left(1-\alpha_B\right)\left(\frac{\overline{K}-K_A\left(L_{A,t-1}+X_{t-1}\right)}{\overline{L}-L_{A,t-1}-X_{t-1}}\right)^{\alpha_B}$$

$$=\left(a_A P_{A,t} + a_B P_{B,t} - \frac{a_A P_{A,t+1} + a_B P_{B,t+1}}{1+R_{t+1}}\right)X_{t-1}$$

The locus of constant variation of employment crosses the locus of constant employment in the steady state. Its slope can be positive or negative, depending on the parameter values.

(iv) The dynamics of the variation of employment. Equations (19) and (20) imply that:

$$\frac{a_A P_{A,t+1} + a_B P_{B,t+1}}{1+R_{t+1}}\Delta X_t = \left(a_A P_{A,t} + a_B P_{B,t} - \frac{a_A P_{A,t+1} + a_B P_{B,t+1}}{1+R_{t+1}}\right)X_{t-1} \tag{21}$$

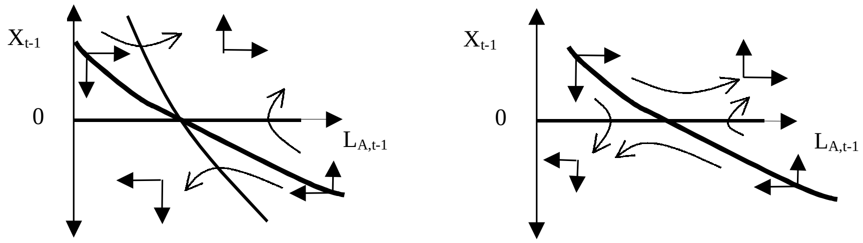
$$-\,P_{A,t}\, H_A\left(1-\alpha_A\right)\left(\frac{K_A\left(L_{A,t}\right)}{L_{A,t}}\right)^{\alpha_A} + P_{B,t}\, H_B\left(1-\alpha_B\right)\left(\frac{\overline{K}-K_A\left(L_{A,t}\right)}{\overline{L}-L_{A,t}}\right)^{\alpha_B}$$

X_t is increasing to the right and decreasing to the left of the locus of constant X_t. Indeed, from (4) and (21):

$$\frac{\partial \Delta X_t}{\partial L_{A,t-1}} = \frac{r_t\left(1+R_{t+1}\right)}{\left(a_A P_{A,t+1} + a_B P_{B,t+1}\right)}\,\frac{\left(1-\alpha_A\right)\ \left(1-\alpha_B\right)\ \left(k_{A,t}-k_{B,t}\right)^2}{\left[\left(1-\alpha_A\right)K_{B,t} + \left(1-\alpha_B\right)K_{A,t}\right]} > 0$$

where $k_{i,\,t}$ denotes capital per capita in sector i. The results in (i) to (iv) determine the phase diagram presented in Figure 1.

Figure 1. The Phase Diagram



Case 1: The locus of constant X has *negative* slope

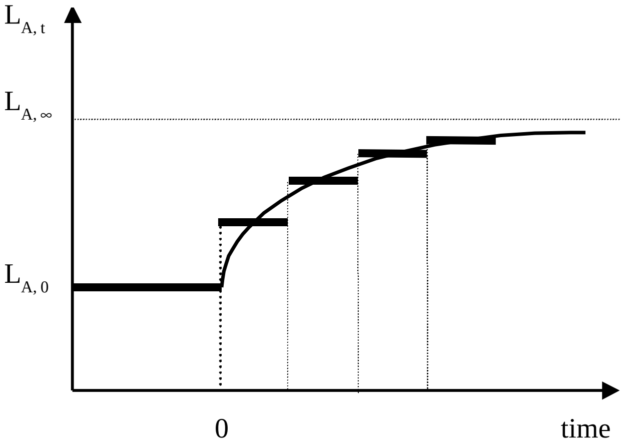
Case 2: The locus of constant X has *positive* slope

The economy exhibits saddle path dynamics. Firms choose how much to increase or decrease employment from the current to the next period ($X_{t-1} = L_{A,\,t} - L_{A,\,t-1}$), given previous period employment $L_{A,\,t-1}$. Rational entrepreneurs pick the value of X_{t-1} on the saddle path, for any other choice would put the economy on an unsustainable path that eventually violates the feasibility conditions of employment $\left(0 \le L_{A,t} \le \bar{L}\right)$.

E. Comparative Dynamics

Consider an increase in the price of sector A that moves the economy away from an initial steady state. The steady state level of employment in sector A rises, and hence both the locus of constant variation of employment ($X_t = X_{t-1}$) and the saddle path shift to the right. Sector A starts hiring new labour. Unlike in the static models, employment does not jump immediately to the new steady state (the new equilibrium in the static model), because of costs of adjustment (see Figure 2). Doing all the adjustment instantly would involve incurring in huge adjustment costs. Rather, entrepreneurs in sector A expand employment gradually, at a pace dictated by the saddle path. Firms in sector B reduce employment at the same velocity firms in sector A expand it, so that total employment remains equal to the exogenous labour supply (see equation (5)).

Figure 2. The Dynamics of Employment in Sector A after an Increase in the Price of Good A



Moving capital is costless in this model. Nevertheless, capital moves gradually from sector B to sector A, at the pace dictated by the movement of labour (equation (17)). Firms in the expanding sector do not want to hire more capital they can efficiently use with the workers they have in each period. Firms in the contracting sector remain using for a while some of the capital they will eventually free. The adjustment costs in one factor determine a slow adjustment not only in that factor but also in other production factors.

The speed of adjustment depends on the adjustment costs in both sectors (equation (19)). The adjustment in sector A is slower the higher is the adjustment cost parameter in sector A, but also in sector B. Firms facing these costs adjust slowly; this is the direct and more obvious effect. But there are also indirect general equilibrium effects going through the returns of production factors that determine a slow adjustment also in the other sector (equations (3)).

The increase in the price of sector A induces a change in the consumption basket. Families reduce consumption of good A relative to good B. Net imports of sector A shrink as production in the sector rises and domestic consumption of this good decreases. Net imports of sector B rise as production reduce and domestic consumption of B increases.

III. Trade Liberalisation, Some Simulation Results

A. Liberalise Now or Later?

Should the government liberalise foreign trade once and for all or should it make the announcement first and give the private sector some time to adjust? There is no point in waiting if, as it is assumed in the standard static HO model of trade, adjusting is costless. But, does this conclusion extend to the more realistic case in which firms do incur in adjustment costs? According to the static HO model, trade liberalisation is good because it induces a more efficient allocation of resources. But, what would be the benefits from trade liberalisation if, because of adjustment costs, resources do not reallocate or do it very slowly? Do adjustment costs provide a rationale for delay or even no liberalisation?

To answer these questions, we compare the general equilibrium welfare effects of eliminating tariffs now or, alternatively, announcing now that tariffs will be eliminated in the future (first two rows in Table 1). Table 1 presents the welfare gains defined as the difference between the sum of discounted utilities with and without trade liberalisation. There is a 15 per cent tariff on the capital intensive import sector in the initial steady state. We consider five values of the adjustment cost parameter, including the limiting case in which the cost of adjustment is zero.

The first conclusion we can draw from Table 1 is that trade liberalisation increases welfare-welfare gains are positive in all these cases. Hence, adjustment costs do not seem to justify keeping positive tariffs, at least not in the scenarios presented in this table. A second conclusion is that liberalising now is better than waiting. Welfare increases more with a sudden immediate tariff elimination than with a postponement and this is so for all the parameter levels considered in these simulations. Welfare gains from a sudden unanticipated trade liberalisation are decreasing in the adjustment parameter (first row in Table 1). Adjustment costs slow down the reallocation of resources and hence reduce the efficiency gains from free trade. In the extreme case of infinite adjustment costs, liberalisation does not induce any reallocation at all.

Nevertheless, small to moderate adjustment costs raise the welfare gains from a pre-announced cut in tariffs (second row in Table 1). Because of

Table 1. Welfare Gains from Trade Liberalisation, Representative Agent Model

	Adjustment cost level				
Timing	Null	Low	Moderate low	Moderate high	High
Unanticipated liberalisation in period 0	516	510	486	411	251
Liberalisation in period 20, announced in period 0	194	197	204	219	170
Liberalisation in period 20, announced in period 20	194	192	183	155	95

adjustment costs, firms start reallocating resources when the government announces that the tariff will be eliminated. Without these costs, firms would not begin the adjustment until the tariff is eliminated. Therefore, the adjustment costs may have a positive effect on economic efficiency after the announcement and before the implementation of the tariff reduction. Adjustment costs still slow down the reallocation of resources after the tariff reduction. These countervailing effects determine that welfare gains from a postponed announced liberalisation are not monotonic in the adjustment parameter.

The effects of the adjustment costs on the welfare gains from trade liberalisation can be interpreted in the light of taxation theory. The larger the tax elasticity of a tax base the larger the welfare losses caused by a distortionary tax, and the larger the welfare gains from eliminating the tax. Adjustment costs reduce the contemporaneous tax elasticity of output, and postpone the efficiency gains from a reduction of a tariff. Hence, the discounted sum of efficiency gains from a sudden and permanent tariff reduction is a decreasing function of these costs. Infinitely large adjustment costs would turn the tariff into a non-distortionary tax. Eliminating the tariff would not contribute to raise efficiency in such a case. But moderate adjustment costs increase the elasticity of current output to a tariff reduction that is known to take place in the future. Therefore, the discounted sum of efficiency gains from a pre-announced liberalisation is an increasing function of the adjustment cost parameter for a range of values.

B. The Value of Pre-announcing Trade Liberalisation

According to the results discussed above, postponing trade liberalisation reduces the welfare gains from this policy. Therefore, there seems to be no room for pre-announcing it. However, real-world changes in trade policy usually take time. Governments seldom eliminate barriers to trade unilaterally. They rather do it after extensive negotiations with other governments. In this more realistic scenario, which are the effects of announcing that barriers to trade will be eliminated in the future? Does the anticipation of tariff reductions increase welfare?

Anticipation of tariff reductions makes future consumption relatively less expensive than current consumption, inducing higher domestic savings and a surplus in the current account of the balance of payments. This phenomenon is the reverse of the well known consumption boom and current account deficit that have been associated to trade liberalisations that are thought to be temporary (Calvo, 1988). The policy implications of this phenomenon in terms of the timing of trade and financial liberalisations have been extensively analysed in the literature (Falvey and Kim, 1992). The productive effects of expected variations in tariffs have been far less analysed.[5] In order to focus on the productive dynamic effects of a pre-announced liberalisation, we get rid off the consumption and savings effects, assuming that the goods are perishable and that the economy has no access to international credit markets. The standard HO model highlights the static productive distortions caused by tariffs. The dynamic version presented in this paper allows for the simultaneous analysis of the static and the dynamic distortions in the allocation of resources.

In principle, good information about economic policy helps private agents to make the right choices. But announcing a tariff reduction adds an inter-temporal distortion to the existing intra-temporal distortion caused by the

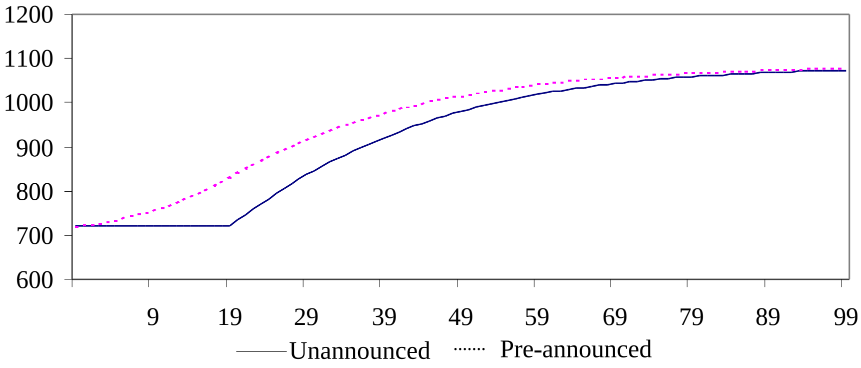
[5] Leamer (1980) analyses these effects in a very simplified two-periods economy.

tariff. The goods affected by the tariff become relatively more expensive not only with respect to other goods in the same period, but also with respect to the same goods in the future. Yet, because of the second-best principle, it is not a-priori obvious whether adding this inter-temporal distortion increases or decreases welfare. To address this issue, we simulated an elimination of the tariff in period twenty, assuming first that agents are informed about this policy in period zero, and assuming later that agents learn about this policy only when the tariff reduction takes place –i.e. agents are surprised–.

The results summarised in Table 1 (rows 2 and 3) indicate that a pre-announced trade liberalisation is more beneficial than a surprise one, i.e. there is a positive value associated with the announcement when there are adjustment costs. Because of them, the reallocation of resources that enhances efficiency begins when the tariff elimination is announced (Figure 3). Therefore, the announcement should not be delayed.

The welfare gains caused by announcing the trade liberalisation –the "value of the announcement"– depend on the adjustment cost parameter. With zero adjustment costs, the information that the tariff will be reduced does not raise welfare. If reallocating resources is costless, firms do not start reallocating productive factors until the tariff is actually reduced, no matter whether they learn about the reduction before or in the very moment in which it takes place. In the simulations reported in Table 1, the "value of the announcement" increases with the adjustment cost parameter. After the announcement and

Figure 3. Employment in the Expanding Sector (Liberalisation in Period 20)



before the tariff is actually eliminated, firms reallocate resources faster the more costly is to do it.

C. Winners and Losers from Trade Liberalisation

Trade would not affect individuals differently if the property rights over productive factors were uniformly distributed in the population or if the government implemented compensating transfers. The representative agent model presented in previous sections assumes that resources are uniformly distributed in the population. This assumption allowed us to focus on the efficiency effects of trade liberalisation, leaving aside the distributional effects of this policy. But the adjustment costs also have some interesting non trivial consequences on the distributional effects of trade liberalisation. In order to address this issue, we consider now a version of the dynamic-HO model with heterogeneous population.

Owners of production factors receive the same return in both sectors, if production factors are not specialised. With non-specialised labour, trade equally affects all workers; the same is true for capitalists. Adjustment costs do not modify this basic property of the HO model. But things are different regarding to the property of firms. Because of adjustment costs, competitive firms make non-zero profits and profits may differ across sectors. While owners of firms in one sector may be making benefits, owners of firms in the other sector may be suffering loses. These considerations led us to identify four distinctive groups in the society: workers, capitalists, owners of firms in sector A and owners of firms in sector B.[6] Of course, societies are usually not so neatly stratified, but this stark assumption about the distribution of property rights is useful to highlight the distributional effects of trade liberalisation. Table 2 summarises the effects of eliminating the tariff in the capital-intensive sector on the welfare of these four different groups.

Workers are among the winners and capitalists are among the losers in this example, because sector B –the one whose tariff is being eliminated– is capital intensive. These are standard results from the static HO model. The

[6] The government is assumed to channel the proceeds of tariffs to consumers of import goods in a lump-sum fashion. This neutral assumption is made to isolate the effects of distortions caused by tariffs from the income extraction effect which is common to any tax.

Table 2. Welfare Gains from Trade Liberalisation, Heterogeneous Population

	Adjustment cost level				
	Null	Low	Moderate low	Moderate high	High
a) Workers					
Unanticipated liberalisation in period 0	2,303	2,271	2,133	1,716	851
Liberalisation in period 20, announced in period 0	869	872	879	886	606
b) Capitalists					
Unanticipated liberalisation in period 0	-1,792	-1,775	-1,693	-1,438	-879
Liberalisation in period 20, announced in period 0	-677	-679	-682	-683	-513
c) Owners of firms in sector A					
Unanticipated liberalisation in period 0	0	13	66	219	515
Liberalisation in period 20, announced in period 0	0	1	2	11	115
d) Owners of firms in sector B					
Unanticipated liberalisation in period 0	0	-4	-24	-88	-235
Liberalisation in period 20, announced in period 0	0	1	2	5	-39

news is that owners of firms in the expanding sector receive a positive discounted sum of profits, while owners of firms in the contracting sector may or may not experience loses. At first glance, the first result looks easier to understand than the second, but more careful analysis shows that both results respond to quite complex general equilibrium dynamic effects. The fact that the elimination of the tariff in sector B “favours” sector A does not imply that firms in this sector must make profits. Depending on the timing of the process, firms in the expanding sector may even experience initial loses (Figure 6 will present an example).

Adjustment costs reduce the impact of a sudden unanticipated trade liberalisation on workers and capitalists (Table 2). The larger the adjustment parameter, the smaller the welfare gains of the former and the welfare loses of the latter. In turn, owners of firms are more affected when reallocating resources is costly: owners of firms in the expanding sector are benefited the more and owners of firms in the contracting sector are damaged the more, the larger the adjustment parameter. Adjustment costs thus shift the burden of the risk of unanticipated trade policy changes from owners of production factors to owners of firms.[7]

Things are more complex in the case of a pre-announced liberalisation. According to the results summarised in Table 2, workers get larger welfare gains and capitalists experience larger loses the larger the adjustment parameter for small and moderate levels. But sufficiently large adjustment costs reduce gains and loses, just as they do in the unanticipated case. The ambiguity stems from the crossing of the return curves for different levels of the parameter (Figures 4 and 5). The wage and the return to capital start to change as soon as the announcement is made. After the policy is announced and before it is implemented, the return to production factors change faster the larger the adjustment parameter. But after the tariff is actually eliminated, the return to production factors change slower the larger are the costs involved. Therefore, in this case adjustment costs do not always reduce the trade policy risk for owners of production factors.

Pre-announcing trade liberalisation has non trivial effects on the value of the firms and the welfare of their owners. The value of the firms in the expanding sector rises in a pre-announced liberalisation, as it does in a surprise unanticipated one. Also, it rises the more, the larger the adjustment cost parameter. But unlike in the unanticipated liberalisation, the value of the firms in the contracting sector may also rise when it is pre-announced, if the parameter is not too large.

[7] It is quite immediate that the same holds true for the risk of variation of international prices.

Figure 4. The Dynamics of the Returns to Capital in a Pre-announced Liberalisation

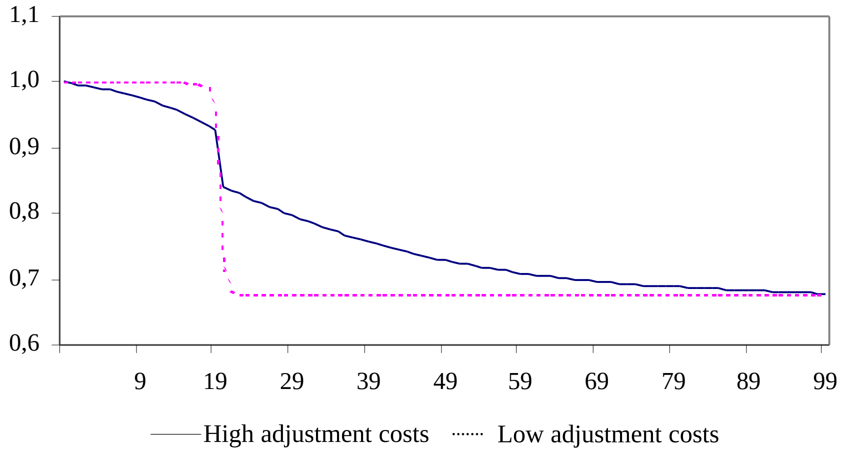
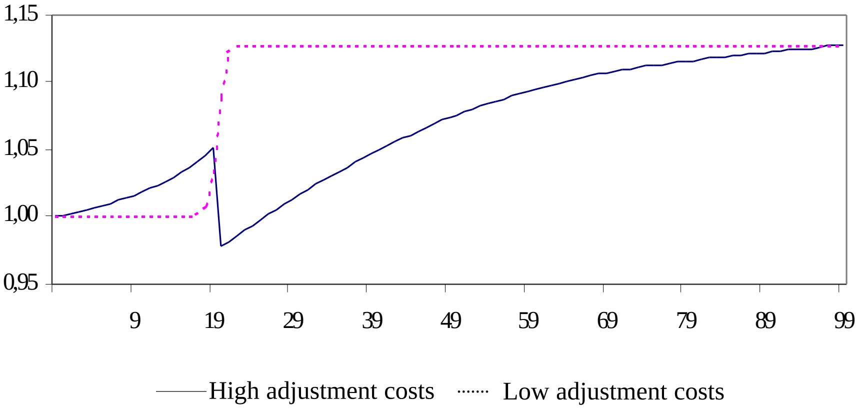
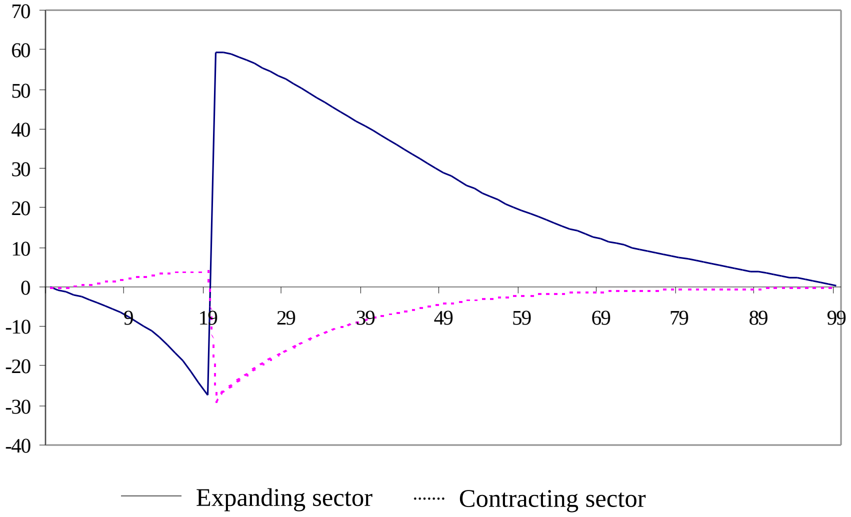


Figure 5. The Dynamics of Wages in a Pre-announced Liberalisation



The possibility that firms in the contracting sector increase their value stems from the depressing effect of the announcement of the tariff elimination on the return to capital, the factor in which the contracting sector is intensive. The news that the protected sector will have to face an output price decline due to the programmed elimination of the tariff, coupled with the existence of costs of adjustment, induces firms in this sector to immediately start firing resources and firms in the other sector to start hiring resources. Being the contracting sector more intensive in the use of capital than the expanding sector, capital becomes relatively abundant while labour becomes relatively scarce. The return to capital decreases and the return to labour increases. The decline in the return to capital relative to the return to labour favours the capital-intensive protected sector and damages the labour-intensive export-oriented sector. Therefore, immediately after the announcement, the expanding sector experiences loses while the other makes profits. When the tariff is eliminated, firms in the formerly protected sector face a sharp one-step decline in the output price and start making loses. Firms in the expanding sector start making profits, as the return to capital drops following the sharp decline in the price of the good in the capital-intensive sector (Figure 6). Because of these complex time profiles of the profits, a pre-announced reduction of a tariff in presence of costs of adjustment may raise the value of the firms even in the sector that is being unprotected. Postponing the measure obviously reduces the present value of the welfare gains and loses caused by the elimination of the tariff. As it comes clear from Table 2, the unanticipated liberalisation in period zero yields larger gains and loses than the liberalisation

Figure 6. Profits in a Pre-announced Liberalisation (High Adjustment Costs)



in period twenty announced in period zero.[8] But this observation is not particularly illuminating: indefinitely postponing the liberalisation would cause no gains and no loses. Not surprisingly, similar conclusions have been reported in quite different frameworks (Mussa, 1986; Albuquerque and Rebelo, 1998).

IV. Concluding Remarks

This paper revisits some of the issues analysed in Mussa (1986), assuming net rather than gross adjustment costs in a dynamic version of a HO model of trade. Some new issues arise. Firstly, as expected, trade liberalisation enhances efficiency and there is no efficiency reason for postponing it in this HO model with adjustment costs. But, if for other reasons, such as distributional concerns and political support, the elimination of tariffs must be postponed, the announcement of the policy has a positive effect on efficiency, speeding up the reallocation of resources. Of course, announcing a future tax reduction may have other distortionary effects on the intertemporal allocation of consumption and savings, making the balance ambiguous. But we make the point that the positive effect of the announcement fostering the reallocation of resources should not be dismissed when reallocating resources is costly. Previous literature on trade liberalisation that has not explicitly considered the costs of adjustment did not take the efficiency value of the announcement into account.

Adjustment costs reduce the efficiency gains from a sudden unanticipated trade liberalisation. This is not surprising since the expected efficiency gains stem from the reallocation of resources that is hindered by costly adjustment. However, small to moderate adjustment costs may *raise* the efficiency gains from a pre-announced liberalisation. Adjustment costs are needed for the announcement of a future elimination of the tariff to induce the reallocation of resources now. With zero adjustment costs, firms would wait until the tariffs are actually eliminated to reallocate resources, and the announcement would be valueless.

These results have implications for the design of reform packages that involve both liberalising foreign trade and removing regulations that slow

[8] The difference is even larger if the liberalisation in period twenty is not pre-announced.

down the reallocation of resources. If the country is engaged in a gradual process of trade liberalisation, it may not be optimal to *fully* remove these regulations until the process of trade liberalisation is complete. Furthermore, it would not be advisable to announce that the regulations that slow down the adjustment process will be removed immediately after the elimination of barriers to trade, for this announcement would eliminate the incentives to reallocate resources before. This result is an application of the second-best principle: removing a distortion may not be beneficial when other distortions remain (for other examples of the same principle, see Edwards, 1988, and Rama, 1997). Unfortunately, this principle is not easily applicable in practice. Imperfect knowledge of the appropriate model and parameter values makes it difficult to determine to what extent regulations that slow down adjustment should be maintained. In any case, this second-best type of argument should be taken into account in any careful assessment of a reform package.

The distributional effects of trade reform in the presence of adjustment costs depend on whether the policy is pre-announced or not. By and large, adjustment costs reduce the welfare gains and loses of owners of production factors from a tariff elimination that is not anticipated. The burden of the risk is mostly shifted to the owners of firms. When adjustment costs are present, pre-announced trade liberalisations have more complex distributional effects than unanticipated liberalisations. Owners of the production factor that is negatively affected by the tariff elimination may experience larger loses with moderate than with low adjustment costs. Owners of firms in the contracting sector may experience welfare gains with a pre-announced liberalisation when adjustment costs are moderate.

The results in this paper suggest that the costs of adjustment matter for the political support for trade liberalisation, but they also suggest that this relationship is complex. On one hand, large adjustment costs dampen the efficiency gains from trade liberalisation and may thus reinforce protectionism. Because of adjustment costs, the efficiency gains from freer trade take time to materialise, reducing the appeal of liberalisation for the government, particularly so if the government has to incur in some short run costs to implement the reform. Moreover, protectionism has often contributed to raise adjustment costs, since non-competitive environments favour lobbying for regulations that create rents and reduce flexibility. Therefore, protectionism and regulations that increase rigidity may reinforce each other in a vicious circle. On the other hand, adjustment costs impact on the distributive effects of trade liberalisation potentially modifying the political support of the reform. Nevertheless, no simple conclusion can be drawn from our analysis in this respect. While some losers from liberalisation experience smaller loses, some other losers suffer larger loses due to the adjustment costs. The opposition to trade reform of the former may be ameliorated, but the opposition of the latter will likely be exacerbated by the costs of adjustment.

The model presented in this paper is a dynamic extension of the standard two-sectors-two-factors HO model of trade. In principle, the same approach could be used to develop a dynamic extension of a HO model with more than two factors and sectors. Such a model would be particularly interesting to analyse the effects of trade liberalisation on the labour skill premium.[9] The increasing skill premium that has accompanied some recent processes of trade liberalisation in developing countries in which unskilled labour is abundant is at odds with the basic predictions of the standard HO model. One possible explanation is, of course, that in these cases the rise in the skill premium does not respond to trade liberalisation, but to technological change or other economic trends. Another complementary explanation could be explored with an extension of the dynamic HO model that included both skilled and unskilled labour. Notice in Figure 5 how the return to the production factor that is eventually benefited with the freeing of trade decreases immediately after the elimination of the tariff in a pre-announced liberalisation, if the adjustment cost parameter is sufficiently large. In this fashion, the return to unskilled labour could well decrease in the initial phase of the liberalisation process and rise later on. The skill premium would thus exhibit a hump shaped path. This is of course just an example, but it does suggest that introducing some relatively simple dynamics can significantly increase the empirical explanatory capacity of the HO model of trade.

[9] The significant rise in wage inequality that has been documented in many countries during the eighties and nineties has received much attention in the literature. Globalisation is one of the competing explanations of this fact. See, among many others, Bound and Johnson, 1992; Acemoglu, 1999; Birdsall and Graham, 2000; and Leamer, 2000.

Appendix. Consumers Program

Adding the consumers per period budget constraints, we can rewrite program (8) with the intertemporal budget constraint:

$$\underset{\{C_{A,t},C_{B,t}\}\ t=0,\ldots,\infty}{Maximise} \quad \sum_{t=0}^{\infty} \beta^t C_{A,t}^{\theta} C_{B,t}^{1-\theta} \tag{A.1}$$

s.t.

$$\sum_{t=0}^{\infty} \frac{P_{A,t}\left(C_{A,t}-Y_{A,t}\right)+P_{B,t}\left(C_{B,t}-Y_{B,t}\right)}{\prod_{i=0}^{t}\left(1+R_i\right)} = A_0$$

We have imposed a transversality condition in the intertemporal budget constraint, namely that the present value of net assets that consumers hold in the infinitely far future is zero:

$$\lim_{t\to\infty} \frac{A_t}{\prod_{i=0}^{t}\left(1+R_i\right)} = 0$$

The first order conditions of this program imply equation (10). Using this result back into (A.1), we rewrite the consumers program as:

$$\underset{\{C_{A,t},C_{B,t}\}\ t=0,\ldots,\infty}{Maximise} \quad \sum_{t=0}^{\infty} \beta^t \left(\frac{1-\theta}{\theta}\right)^{1-\theta} \left(\frac{P_{A,t}}{P_{B,t}}\right) C_{A,t} \tag{A.2}$$

s.t.

$$\sum_{t=0}^{\infty} \frac{\left(1/\theta\right)P_{A,t}C_{A,t}-\left(P_{A,t}Y_{A,t}+P_{B,t}Y_{B,t}\right)}{\prod_{i=0}^{t}\left(1+R_i\right)} = A_0$$

$$C_{B,t} = \left(\frac{1-\theta}{\theta}\right)\left(\frac{P_{A,t}}{P_{B,t}}\right)C_{A,t}$$

This is a linear programming problem. Indifference curves and budget lines in the *($C_{A,\,t}$, $C_{A,\,t+1}$)* space are both straight lines. The program yields corner solutions unless the slope of the budget lines and the indifference curves coincide, in which case consumers are indifferent between consuming in t or in t + 1. Corner solutions are not consistent with credit market equilibrium, so these slopes must coincide:

$$\left.\frac{dC_{A,t+1}}{dC_{A,t}}\right|_{budget} = -\frac{P_{A,t}}{P_{A,t+1}}\left(1+R_{t+1}\right) \tag{A.3}$$

$$= \left.\frac{dC_{A,t+1}}{dC_{A,t}}\right|_{indifference} = -\frac{1}{\beta}\left(\frac{P_{A,t}}{P_{A,t+1}}\frac{P_{B,t+1}}{P_{B,t}}\right)^{\theta}$$

Equation (9) follows.

References

Acemoglu, D. (1999), "Patterns of Skill Premia," *NBER Working Paper Series 7018*.

Albuquerque, R., and S. Rebelo (1998), "On the Dynamics of Trade Reform," *NBER Working Paper Series 6700*.

Birdsall, N., and C. Graham, eds. (2000), "New Markets, New Opportunities? Economic and Social Mobility in a Changing World," *Brookings Institution Press*, Washington, D.C.

Bound, J., and G. Johnson (1992), "Changes in the Structure of Wages in the 1980's: An Evaluation of Alternative Explanations," *American Economic Review* **82**: 371-392.

Calvo, G. (1988), "Costly Trade Liberalisation: Durable Goods and Capital Mobility," *IMF Staff Papers* **35**: 461-473.

Calvo, G., and E. Mendoza (1994), "Trade Reforms of Uncertain Duration and Real Uncertainty: A First Approximation," *IMF Staff Papers* 41**:** 555-586.

Choksi, A., and D. Papageorgiou (1986), "Economic Liberalization: What Have We Learned," in A. Choksi and D. Papageorgiou, eds., *Economic Liberalization in Developing Countries*, 1-11, Basil Blackwell, UK.

Clarete, R., I. Trela, and J. Whalley (1994), "Evaluating Labour Adjustment Costs from Trade Shocks: Ilustrations for the U.S. Economy Using an Applied General Equilibrium Model with Transactions Costs," *NBER Working Paper 4628*.

Dixit, A. (1992), "Investment and Hysteresis," *Journal of Economic Perspectives* **6** (1): 107-132.

Edwards, S. (1988), "Terms of Trade, Tariffs, and Labour Market Adjustment in Developing Countries," *World Bank Economic Review* **2** (2): 165-185.

Edwards, S. (1989), "On the Sequencing of Structural Reforms," *NBER Working Paper 3138*.

Edwards, S. (1994), "Trade and Industrial Policy Reform in Latin America," *NBER Working Paper 4772*.

Falvey, R., and C.D. Kim (1992), "Timing and Sequencing Issues in Trade Liberalisation," *The Economic Journal* **102**: 908-924.

Funke, N. (1993), "Timing and Sequencing of Reforms: Competing Views and the Role of Credibility," *Kyklos* **46** (3): 337-362.

Grossman, G. (1983), "Partially Mobile Capital. A General Approach to two Sector Trade Theory," *Journal of International Economics* **15**: 1-17.

Hamermesh, D. (1989), "Labour Demand and the Structure of Adjustment Costs," *American Economic Review* **79**: 674-689.

Hamermesh, D. (1993), "Labour Demand and the Source of Adjustment Costs," *NBER Working Paper 4394*.

Hamermesh, D., Hassink, W., and J. van Ours (1994), "New Facts about Factor-Demand Dynamics: Employment, Jobs and Workers," *NBER Working Paper Series 4625*.

Karp, L., and T. Paul (1994), "Phasing in and Phasing out Protectionism with Costly Adjustment of Labour," *The Economic Journal* **104**: 379-1392.

Kemp, M., and H. Wan (1974), "Hysteresis of Long-run Equilibrium from Realistic Adjustment Costs," in G. Horwich and P. Samuelson, eds., *Trade, Stability and Macroeconomics*, New York Academic Press.

Leamer, E. (1980), "Welfare Computation and the Optimal Staging of Tariff

Reductions in Models with Adjustment Costs," *Journal of International Economics* **10**: 21-36.

Leamer, E. (2000), "What's the Use of Factor Contents?" *Journal of International Economics* **50**: 17-49.

Michaely, M. (1986), "The Timing and Sequencing of a Trade Liberalization Policy," in A. Choksi and D. Papageorgiou, eds., *Economic Liberalization in Developing Countries*, 41-67, Basil Blackwell, UK.

Mussa, M. (1986), "The Adjustment Process and the Timing of Trade Liberalisation," in A. Choksi and D. Papageorgiou, eds., *Economic Liberalization in Developing Countries*, 68-124, Basil Blackwell, UK.

Neary, P. (1978), "Short-run Capital Specificity and the Pure Theory of International Trade," *The Economic Journal* **88**: 488-510

Neary, P. (1982), "Intersectoral Capital Mobility, Wage Stickiness and the Case for Adjustment Assistance," in J. Bhagwati, ed., *Import Competition and Response*, Chicago University Press.

Oi, W. (1962), "Labour as a Quasi-fixed Factor," *Journal of Political Economy* **70** (6): 538-55.

Papageorgiou, D., Michaely, M., and A. Choski, eds., (1991), *Liberalizing Foreign Trade*, Volumes 1-7, Cambridge, Mass. and Oxford: Basil Blackwell.

Rama, M. (1997), "Labour Market Institutions and the Second-best Tariff," *Scandinavian Journal of Economics* **99** (2): 299-314.

Rothschild, M. (1971), "On the Cost of Adjustment," *Quarterly Journal of Economics* **85**: 605-622.

Sargent, T. (1978), "Estimation of Dynamic Labour Demand Schedules under Rational Expectations," *Journal of Political Economy* **86** (6): 1009-1044.

Journal of Applied Economics, Vol. VI, No. 1 (May 2003), 127-156

WHY IS FRENCH EQUILIBRIUM UNEMPLOYMENT SO HIGH? AN ESTIMATION OF THE WS-PS MODEL

Yannick L'Horty[*]
EPEE, University of Evry Val d'Essonne
and
Christophe Rault[*]
EPEE and EUREQua, University of Paris I Panthéon-Sorbonne

Submitted July 2000; accepted April 2002

Unemployment in France rose steadily from the early-seventies to the mid-eighties. Since the mid-eighties it has continued to experience fluctuations around a very high average level. Equilibrium unemployment theories are a useful framework within which to account for these developments. A multivariate estimation of the WS-PS model on macroeconomic quarterly data, which includes a larger number of potential unemployment determinants than earlier work, allows an enriched reading of the rise in French unemployment and of its persistence at a high level. We estimated it using a conditional VAR-ECM model, which is based upon the weak exogeneity properties of variables over the 1970-1/1996-4 period. The rise in equilibrium unemployment by 10 points in 25 years can essentially be explained by the rise in tax and social wedge, the slowdown in labour productivity and the deterioration of job security. Terms of exchange and skill mismatch account for only a slim part of the rise in equilibrium unemployment.

JEL classification codes: C32, E24
Key words: labour market, WS-PS model, equilibrium unemployment, co-integration, conditional VAR-ECM model

I. Introduction

In France, the unemployment rate has hovered at around 10% for over 20

[*] Rault (corresponding author): chrault@hotmail.com; http://www.multimania.com/chrault/index.html. Maison des Sciences de l'Economie, 106-112 boulevard de l'Hôpital, 75647 Paris, Cedex 13, France. L'Horty: lhorty@eco.univ-evry.fr.

years. Whilst it has experienced significant fluctuations, it has always moved back to this average level. That is why special attention has been granted to structural unemployment theories and their empirical evaluation, in France, as well as in other European countries with comparable evolutions.

Equilibrium unemployment theoreticians commonly substitute a structural relation called WS for Wage Schedule (Lindbeck, 1993), for the labour supply from households in the traditional equilibrium of the labour market. The shape of this relation is deduced from theoretical models most often based on the microeconomic behaviours described by the new labour market theories (e.g. efficiency wages, bargaining models, insider/outsider approach). This relation intersects with another one describing structural price setting (PS). They jointly determine the equilibrium unemployment level that will be modified by structural shocks affecting the determinants of wage or price setting, notably oil crises, shocks on the level of direct or indirect taxes and real interest rate shocks. This sensitivity to structural shocks differentiates the approaches in terms of equilibrium unemployment, qualified as structuralism by Phelps (1994), from those in terms of natural unemployment, as defined by Friedman (1968). Moreover, it leads to a higher unemployment determinant set than the one usually considered by a Phillips' curve approach (Bean,1994). The theoretical WS-PS models have been popularised through the work of Layard, Nickell and Jackman (1991). They have now integrated employed worker heterogeneity (for an example, see Laffargue, 1995) and the dynamic aspects of wage setting (Manning, 1993; Cahuc and Zylberberg, 1998). This theoretical maturity has resulted in an impressive extension in the list of potential unemployment explanations, which both rest on an explicit microeconomic base and are connected to wage or price schedules in a general equilibrium framework.

This theoretical maturity contrasts with the state of empirical research whose purpose is to estimate the WS-PS model. The literature on this topic can be categorised into two separate groups. The univariate estimations of the WS and PS relations are compatible with a large number of unemployment equilibrium determinants, in accordance with the theory, but do not take the interdependences between variables into account. Inversely, too large a number of variables become incompatible in practice with a multivariate estimation of the WS and PS relations, yet it is more satisfactory to take the interdependences between wage and price setting into account. In French macroeconomic data, the equilibrium unemployment rise since the early seventies was thus entirely explained by real interest rate evolution, technical progress and the terms of exchange in Bonnet and Mahfouz (1996), by the evolution of the wage wedge, the replacement ratio and productivity in L'Horty and Sobczak (1997), and by the evolution of capital cost and the wage wedge in Cotis, Méary and Sobczak (1997). These multivariate estimations put the emphasis on the crucial role of some variables but do not fully explain the rise and persistence in unemployment.

The contributions of this paper are essentially threefold. To begin, it focuses on a structured theoretical setting that deals with a large number of potential equilibrium unemployment candidates. It then proceeds to build a set of original indicators for some of these determinants. Finally, it uses an econometric methodology to consider the effects of these variables simultaneously. This allows a greater understanding of the formation of equilibrium unemployment than that of the existing applied studies on French data. This reading is theoretically justified, compatible with the statistical properties of the variables considered, and validated by multivariate econometric techniques, which leads to a retrospective and quantitative explanation of French unemployment over the 1970/1-1996/4 period.

As to the econometric methodology, this paper gives an estimation of the WS-PS model on French macroeconomic data that is both in keeping with Johansen's multivariate estimation techniques and compatible with a large number of variables.[1] This re-estimation is made possible by taking the weak exogeneity properties of variables into account. The multivariate model can indeed be partitioned in two blocs whose parameters vary freely: a marginal model gathering the weakly exogenous variables for the long run parameters

[1] Our estimation is purely national and enables estimations obtained to be completed with multinational data using panel econometric techniques (cf. for example Layard, Nickell, Jackman, 1991 and Layard, Nickell, 2000). A comparative approach on international data imposes great restrictions in the construction of data that must be homogeneous between countries. In a purely national study, we do not have this constraint of data homogeneity, which allows us to construct more representative indicators of the French situation. This is, for example, the case for a complete set of SMIC hikes, replacement ratio, working hours, or progression of social wedge. These data would be either impossible to build for other countries or feebly representative of the French situation in an internationally standardised database.

of the Vector Error Correction Model (VAR-ECM), and a conditional model composed of other equations. Co-integrating vectors can then be estimated from only the conditional model, reducing the system size without losing any information from the full VAR-ECM.

Starting from a quarterly database composed of 16 series and covering the 1970/1-1996/4 period, we estimated the WS-PS model using an unrestricted VAR-ECM approach, composed of ten variables. Two co-integration relations were estimated from a partial system composed of seven equations and conditional to the three equations describing the evolution of weakly exogenous variables. These relations were identified using an approach inspired by Manning (1993), according to which productivity is not in the structural wage equation. It is important to note that the equilibrium unemployment estimation is robust with respect to that identification constraint. Finally, exclusion tests retained only five determinants in the progression of unemployment equilibrium in France: hourly productivity, through which real interest rates can have an impact; the internal terms of exchange, which essentially vary under the impact of oil crises and the exchange rate; the quit ratio; the aggregate wage wedge through which the different deduction rates can have an influence; and skill mismatch. The method used allows a calculation of the respective influences of these determinants and their retrospective contributions to unemployment development. On the other hand, the replacement ratio, which depends on the generosity of the unemployment benefit system, working hours, the French minimum wage (*SMIC*) increase and the progressiveness of the social wedge would have had a non-significant role in the evolution of equilibrium unemployment according to this estimation.

Section II provides a theoretical review of the WS-PS model. It presents the list of potential variables that can account for unemployment equilibrium, the mechanisms through which these variables have an influence and the data used in this study, which required that several original indicators be constructed for the different variables. Section III presents the model estimation results. Finally, Section IV presents our conclusions.

II. Equilibrium Unemployment Determinants and their Measures

Ideally, the richest possible theoretical model would stem from a microeconomic wage and price setting base in a dynamic framework that would take agent anticipation setting into account, as well as nominal and real rigidities and the impact of labour market institutions, such as systems of employment protection, trade union activity, active labour market policy, and so forth. Such a labour model would contain a heterogeneous factor, where all deductions and transfer systems would be modelled, including the modalities of unemployment benefit payments, their digressiveness in time and more generally, the degree of progress in the fiscal and social system. On that basis, one would deduce both a short and long term structural form of WS and PS in a general equilibrium framework to describe all determinants of equilibrium unemployment. Given all these enrichments, there is most probably no analytic solution as to the log-linearisation of structural wage and price curves. Moreover, the specification of log-non-linear structural expressions of these curves would be highly dependent on the whole successive modelling choices, and would make a non-linear estimation very delicate. In any case, writing such a full model seems impossible.

Consequently, the estimation strategy adopted here is less ambitious. From the theory, we have selected a list of variables, their expected signs, possibly some bounds for their elasticities and no more. We can then let data speak for themselves in a multivariate log-linear estimation framework.

A. Theoretical Variables

A first list of variables is given by a WS-PS model inspired by Layard, Nickell and Jackman (1991). In that model, goods markets are in imperfect competition and wages are the result of a negotiation between unions and employers, the latter maintaining their right to manage. This static homogenous labour factor model is what enables us to describe the traditional determinants of price and wage schedule and equilibrium unemployment.

In a formal definition of the value of unemployment equilibrium, one solves the system composed of the WS and PS structural equations by substituting the wage share in the added value when a Cobb-Douglas technology is used. One thus obtains a reduced form of the wage equation that defines the level of equilibrium unemployment. In the Layard, Nickell and Jackman (1991) model, this reduced form is presented as the structural form of WS. Equilibrium unemployment increases, ceteris paribus, with union power, the replacement ratio and employees' risk aversion. It decreases with the risk of becoming unemployed, with the degree of competition on the goods market, and with the labour factor efficiency parameter. It is also sensitive to the terms of exchange and to all the parameters characterising the tax system, which play a role in the wage wedge and modify the replacement ratio.

In the case of a CES production function, the structural wage equation remains the same, but it is no longer the case for the equilibrium unemployment expression, which in addition now has a productivity term whose impact depends on the substitution elasticity of factors. If factors are less substitutable than in the case of a Cobb-Douglas technology, the equilibrium unemployment elasticity to labour productivity in efficiency units is negative. An increase in productivity leads to both a wage increase and an unemployment decrease. If factors are more substitutable than in the case of a Cobb-Douglas, productivity in efficiency units has a positive impact on equilibrium unemployment. In other respects, technical progress can be seen to have no impact on equilibrium unemployment levels and to lead only to a real wage increase.

The interest rate influence goes through the productivity term. In the case of a Cobb-Douglas technology, an increase in interest rates reduces equilibrium capital intensity, decreases labour productivity, increases equilibrium labour costs and finally increases equilibrium unemployment. An increase in real interest rate always leads to a decrease in productivity, but it yields to a decrease in equilibrium unemployment if factors are more substitutable than in the case of a Cobb-Douglas technology, and to an increase in the opposite case (PS variations more than compensate those of WS in the former case). This result is not non-intuitive: when factors are slightly substitutable, a capital cost increase limits the use of all factors and thus increases equilibrium unemployment; when they are very substitutable, the substitution effect is bigger than the income effect and equilibrium employment increases.

This model can be completed by specification enrichments, which introduce new variables, by taking into account the dynamic aspects of wage and price schedules and by the introduction of labour heterogeneity. A first specification enrichment consists of introducing working hours. If hours and men are perfect substitutes concerning the technology used by firms, and if a reduction in working hours is not compensated by a rise in hourly wages, taking working hours into account would not change the PS expression. A reduction in working hours can also affect wage setting, according to the individual and union utility functions and the way this reduction is implemented (imposed or bargained). Another specification enrichment is in no longer assuming that the different deductions are flat. Then, if the progressiveness of social or fiscal deductions is taken into account, the price equation remains unchanged but wage equation is distorted, a stronger progressiveness having the same effect as a reduction of union market power in the bargaining. Moreover, in the Layard, Nickell and Jackman model (1991), a φ parameter is introduced to weight unemployment rates in the expression of the employed workers' withdrawal in the bargaining. This parameter represents the risk of becoming unemployed as a function of unemployment rate. Unemployment risk can also be measured in reference to the short length unemployment rate or to the quit ratio extracted from data flows on the labour market. This latter extension is also essential when the dynamic aspects of wage setting are taken into account. Finally, taking employed worker heterogeneity into account leads to other enrichments in the understanding of employment setting. If one distinguishes between different qualifications, one takes the consequences of the skill mismatch on the labour market into account.

All in all, the initial theoretical model and its enrichments lead the price and wage schedule to depend on apparent labour productivity or on the real interest rate, on the price-elasticity of demand, on the efficiency of the labour factor (which corresponds in a Cobb-Douglas production function to the share of wages in added value) and on working hours. As far as real wage setting is concerned, it depends on the unemployment rate, on union bargaining power, on the degree of competition in the goods market, on employed workers' risk aversion, on the replacement ratio, on the wage wedge and its components, on working hours, on wage wedge progressiveness, on the quit ratio and on the skill mismatch. Equilibrium unemployment depends on all these determinants as soon as their elasticities differ in the price and wage equations.

B. Indicators for those Variables

The empirical evaluation of equilibrium unemployment is faced with a data deficit. Some determinants of the WS-PS models are not directly observable and cannot therefore be found in any existing database. This is the case of price elasticity for goods demand, which embodies the degree of competition between offers on the product markets. It is also the case of the mark-up battle between employed workers' and employers' representatives in wage bargaining, of employed workers' risk aversion or of their psychological discount rate. Other theoretical determinants of equilibrium unemployment can be observed in a more or less direct way, but are not the subject of standardised statistic series (as is true in the case of replacement ratio or of wage wedge progressiveness, for instance). Given this data deficit problem, one answer is to build indicators for these variables. The asset of building indicators is to produce new statistics containing information on market labour evolution.

Most traditional data consists of gross wages, prices, added value, and rates of unemployment. We used the average gross hourly wage rate in the non-financial non-agricultural manufacturing sectors, which was extracted from quarterly accounts. This is also the case for consumption prices, and for added value prices and employment, which were all re-calculated for the non-financial non-agricultural manufacturing sectors. Two apparent labour productivity indicators were used: productivity per capita, which is the ratio of added value to employed workers, and hourly productivity, which is the ratio of per capita productivity to working hours.

Working hours are the synthetic indicator calculated by the French Ministry of Labour. It takes part-time job development into account, which has been promoted over the recent period by state specific assistance (a basic reduction of social wedge to share part-time jobs, some modalities of social contribution reduction on low wages that were encouraging part-time jobs). This indicator dropped throughout the nineties, falling more sharply after 1993, because of the accelerated diffusion of part-time jobs. This indicator is closer to the average working hours really performed by workers.

Real interest rate is the price of public and semi-public bonds. Its direct introduction into a price equation justifies itself when one considers the capital setting as endogenous and when one considers the existence of an asymmetry in capital and labour mobility. In the case of a small open economy in a perfectly integrated worldwide capital market, the interest rate is fixed from abroad and involves capital intensity and equilibrium productivity, which is decisive for price behaviour. An increase in interest rates reduces equilibrium capital intensity, which leads to a decrease in equilibrium labour costs and to a rise in unemployment (PS is horizontal and moves downwards).

The global wage wedge is composed of the internal terms of exchange, which are the ratio of consumption prices to producer prices, and of the social and fiscal wedge, which is itself composed of the social wedge (employers' and employees' contribution rates) and of the fiscal wedge (*VAT*, income tax rate). Employers' and employees' contribution rates (*CSE* and *CSS*) are extracted from social scales, applied to medium wage and given the same progression as the social security ceiling. Direct or indirect (Personal Income Tax and *VAT*) income tax rates, are taken from the databases of the French Ministry of Finances. Theoretically, only the deductions that are not considered by employed workers as benefits or postponed income compensations exert an upward pressure on labour cost and equilibrium unemployment.

For the replacement ratio, we used the indicator created by the Unédic (1997), which is an average of the situations of all unemployed workers at a given date. An extension of unemployment duration leads to a replacement rate reduction, which provides a satisfactory result. This quarterly indicator has been available since 1986. For previous years, we used the unemployment benefit scales applied to the situation of a medium unemployed worker whose period out of work is given by long series employment surveys (we also assumed a 6-12 month affiliation duration). Spontaneously, the two series were very close in 1986. The replacement ratio was clearly on the decrease after the 1992 reform of unemployment benefits.

To measure the quit ratio, which includes the risk of losing one's job and can be linked with the systems of labour protection, we used the transition rate between employment and unemployment, extracted from employment survey, and made it quarterly by a simple linear interpolation. It is important to notice that this rate is not directly connected to the unemployment rate: more intensive flows from employment to unemployment do not imply an increase of unemployment rate, since transitions from inactivity can decrease and exit employment rate can rise. Inversely, an employment flow reduction to unemployment does not imply an unemployment decrease, since these flows can be compensated by an increase of the transitions from inactivity to unemployment, or by a reduction of unemployment exits to employment or inactivity. This transition rate from employment to unemployment is an approximate measure of the probability of being laid off, which can vary in an inverse way to unemployment rate.

Employed workers' bargaining power is one of the parameters on which we have very little information. Instead of using a simple trend or a unionisation rate, whose reading is complex in the case of France, we have used the complete set of hikes given to the minimum wage (*SMIC*). It is an indirect proxy, whose justification is less to demonstrate the wage scale rigidity when the *SMIC* is increased, than to synthetically sum up the evolution of the general climate around wage setting.

The progressiveness of the wage wedge (*PROG*) is calculated here using the residual progressiveness indicator proposed by Jakobsson (1976). The progressiveness of the contributions of employers and employees are calculated separately and the aggregate indicator is obtained by summation.

The mismatch indicator (*MM*) is the semi-variance of relative employment rates by qualification, whose theoretical reading is given by Jackman, Layard and Savouri (1991): when wage curves are convex, a greater dispersal of unemployment rates induces an upward pressure on wages, which leads to a higher equilibrium unemployment rate. Sneessens's indicator (1994) is also tested. It deals with the ratio of the share of qualified employed workers in employment to their share in the labour force.

Other institutional variables could be taken into account when dealing with international approaches using panel data estimation techniques. Thus, centralism of wage bargaining, the systems of labour protection (for the part that does not affect the quit ratio) and active labour market policy can influence wages and unemployment formation. Without any time series data available for these variables, these determinants will be included in our econometric estimation by the constant, or, if they have varied across time, by the trend of our wage and price equations.

III. WS-PS Model Estimation

This section describes the statistical properties of the series as well as the results of the unrestricted VAR-ECM modelling that we finally adopted.

A. Univariate Properties of the Series

The database is composed of 15 quarterly series. It concerns the non-agricultural manufacturing sector and covers the 1970-1 to 1996-4 period. Deduction rates can be regrouped in two levels of aggregation, adding four indicators more.

The first step in the analysis was simply to look at the data univariate properties and to determine the degree to which they were integrated. Theoretically, a process is either I(0), I(1) or I(2). Nevertheless, in practice, many variables or variable combinations are borderline cases, so that distinguishing between a strongly autoregressive I(0) or I(1) process (interest rates are a typical example), or between a strongly autoregressive I(1) or I(2) process (nominal prices are a typical example) is far from easy. We therefore applied sequences of standard unit root tests, i.e. the augmented Dickey Fuller tests, namely the Jobert, 1992, procedure, as well as the Schmidt and Phillips, 1992, test and the Kwiatkowsky, Phillips and Shin (KPSS), 1992, test, to investigate which of the I(0), I(1), I(2) assumptions is most likely to hold true. The results of the Jobert procedure, Schmidt and Phillips' test and the KPSS tests are shown in Table 1. Note that all variables were transformed in natural logarithm, and in what follows lower-case letters denote the natural logarithm of the corresponding variable.Most variables seemed well characterised as an I(1) process, some with non-zero drift. Nevertheless, concerning *u*, *cp*, *pc-p* and *tr*, the results given by the different tests were not all concomitant and did not allow us to decide between an I(0) or I(1) process: they diverged on the number of lags to introduce to have white noise residuals, and on the applied unit root test. The fact that real wages were I(1) supported the estimation of a real model. While considering wages and prices separately, one was likely to introduce variables I(2) in estimations that would not be compatible with the econometric methodology adopted here. Moreover, this would strongly complicate the partition between marginal and conditional models and would not consequently permit us to provide an enriched reading of unemployment formation. Besides, econometric estimations available in France highlight the unit indexing of wages on prices,

Table 1. Unit Root Test Results

Non-agricultural	Jobert Tests			Schmidt-Phillips Tests			KPSS Tests		
manufacturing sectors	Bic	Hannan	Kmax	Bic	Hann.	Kmax	0	4	8
w - *p*: real labour cost	I(1)	I(1)	I(1)	I(1)+T	I(1)+T	I(1)+T	2.52 I(1)	0.53 I(1)	0.31 I(1)
prodh: hourly productivity	I(1)+T	I(1)+T	I(1)+T	I(1)+T	I(1)+T	I(1)+T	2.16 I(1)	0.49 I(1)	0.30 I(1)
tr: replacement rate	I(1)	I(1)	I(1)	I(1)	I(1)	I(1)	1.38 I(1)	0.30 I(1)	0.19 I(1)
cp: complete set of SMIC hikes	I(1)	I(0)+C	I(1)+T	I(1)	I(0)+T	I(1)+T	2.16 I(1)	0.46 I(1)	0.28 I(1)
r: real interest rate	I(1)	I(1)	I(1)	I(1)	I(1)	I(1)	0.82 I(1)	0.20 I(1)	0.14 (?)
ec: quit ratio	I(1)	I(1)	I(1)	I(1)	I(1)+T	I(1)+T	1.92 I(1)	0.43 I(1)	0.27 I(1)
mm: mismatch	I(1)	I(1)	I(1)	I(1)	I(1)	I(1)	1.87 I(1)	0.40 I(1)	0.24 I(1)
u: unemployment rate	I(0)	I(0)	I(1)	I(1)+T	I(1)+T	I(1)+T	2.30 I(1)	0.48 I(1)	0.28 I(1)
h: working hours	I(1)	I(1)	I(1)	I(1)+T	I(1)+T	I(1)+T	2.26 I(1)	0.49 I(1)	0.29 I(1)
coin: global wage wedge	I(1)	I(1)	I(1)	I(1)+T	I(1)+T	I(1)+T	1.89 I(1)	0.41 I(1)	0.24 I(1)
pc - *p*: terms of exchange	I(0)	I(1)	I(0)	I(1)	I(1)	I(1)	1.32 I(1)	0.30 I(1)	0.18 I(1)
coinfs: fiscal and social wedge	I(1)	I(1)	I(1)	I(1)+T	I(1)+T	I(1)+T	1.80 I(1)	0.41 I(1)	0.26 I(1)
coins: social wedge	I(1)	I(1)	I(1)	I(1)+T	I(1)+T	I(1)+T	2.04 I(1)	0.48 I(1)	0.29 I(1)

Table 1. (Continued) Unit Root Test Results

Non-agricultural	Jobert Tests			Schmidt-Phillips Tests			KPSS Tests		
manufacturing	Bic	Hannan	Kmax	Bic	Hann.	Kmax	0	4	8
coinf: fiscal wedge	I(1)+C	I(1)+C	I(1)	I(1)	I(1)	I(1)	0.82 I(1)	0.21 I(1)	1.49 I(1)
css: employees' social contributions rate	I(1)	I(1)	I(1)	I(1)	I(1)	I(1)+T	2.16 I(1)	0.50 I(1)	0.30 I(1)
cse: employers' social contributions rate	I(1)	I(1)	I(1)	I(1)+T	I(1)+T	I(1)+T	1.00 I(1)	0.24 I(1)	0.15 I(1)
vat: value added tax	I(1)	I(1)	I(1)	I(1)	I(1)	I(1)+T	0.31 I(1)	0.08 I(0)+T	0.06 I(0)+T
tir: income tax rate	I(1)	I(1)	I(1)	I(1)+T	I(1)+T	I(1)+T	1.16 I(1)	0.25 I(1)	0.17 I(1)
prog: progressiveness of social wedge	I(1)	I(1)	I(1)	I(1)	I(1)	I(1)	1.53 I(1)	0.34 I(1)	0.21 I(1)

Notes: a) All series are in logarithm and all the tests were programmed with the GAUSS software. b) Unlike the Jobert and Schmidt-Phillips unit root tests, the null hypothesis of the KPSS test is here the deterministic non-stationarity around a linear trend against the alternative hypothesis of stochastic non-stationarity (presence of a unit root). The critical value at a 5% level is 0.463. c) The question mark "?" in some boxes indicates the difficulty in concluding between an I(0) or I(1), given that the computed test is too close to the 5% critical value.

which also justified the choice of a real model. Therefore, nominal rigidities would not explain unemployment in the long-term horizon that is ours.[2]

B. Estimation Strategy

Given that most of the series in our database are non-stationary trending variables, our analysis is conducted within a framework that allows both for non-stationary and potentially co-integrated variables. Our econometric procedure is close to the multivariate co-integrated systems analysis developed originally by Johansen (1988), then expanded and applied in Johansen (1995). It consists of full information maximum likelihood estimation (FIML) of a system characterised by r co-integrating vectors (CIVs). Under conventional hypotheses the statistical model is the following (see Rault 1997 for a detailed presentation):

$$\Delta X_t = \sum_{i=1}^{P-1} \Gamma_i \, \Delta X_{t-i} + \alpha\beta' X_{t-1} + \Phi D_t + \varepsilon_t, \qquad t = 1,..,T \qquad (1)$$

where (X_t), t = 1,...,T, is a dimensional vector process composed of stochastic variables, $\varepsilon_t \sim$ iid, N $(0_n, \Sigma)$, Γ_i, i = 1,...p-1 are (n, n) matrices, supposed constant in time, α and β are (n, r) non-singular matrices of rank $0 < r < n$, D_t is a vector of non-stochastic variables (constant drift, linear deterministic trend, ...), and Σ is a regular, positive define variance-covariance matrix.

The co-integrating vectors are the β_j columns of the β matrix. In particular, the $\beta_j' X_t$ (j = 1,.., r) can be regarded as stationary linear combinations of non-stationary variables and the α as the weights of these different combinations in each equation of the model.

Then, once the number of co-integrating vectors was determined it seemed natural to more precisely apprehend the structure of the adjustment space, spanned by the α. Applying a test on α, boils down to asking oneself if the long run relation(s) belongs to all the model equations. It deals with a weak

[2] An alternative coherent approach with nominal rigidities supposes the consideration of a modelling of variables in growth rates and not in level. This leads to an estimate of a Philips curve and not a wage curve. For an example of that estimation strategy on French data, cf. Heyer, Le Bihan and Lerais (2000).

exogeneity test of the different variables of the system for long run parameters, whose aim is to check if the sufficient condition given by Johansen (1992) checks out empirically. According to Johansen, if the (X_t) variables of the system are divided into (Y_t, Z_t), a sufficient condition for a variable (or a group of variables) Z_t to be weakly exogenous for long run parameters is that the co-integrating vectors do not belong to the model equation(s) describing the evolution of ΔZ_t. In this case, the joint density function can be factorised into two blocs whose parameters vary freely: a ΔZ_t marginal model gathering the weakly exogenous variables for the long run parameters of the VAR-ECM model, and a conditional ΔY_t model composed of the other equations. The co-integration vectors can then be estimated only from the conditional model, which enables the size of the system to be reduced without losing any information from the full VAR-ECM.[3]

Finally, once the co-integrating relationships had been identified (see Johansen and Juselius, 1994 for a detailed presentation), particular structural hypotheses on the α and β matrices could be tested using asymptotically chi-squared distributed test statistics.

C. Estimation Results

Before choosing the final model, we made much prior estimation, whose main results we can only summarise. Firstly, it was impossible to estimate a satisfactory model when the complete set of SMIC hikes and progressiveness indicators were taken into account. Moreover, it was impossible to get a satisfactory estimation when the Sneessens (1994) indicator was introduced and the estimations were made using the Jackman, Layard and Savouri (1991) indicator, which was significantly different from zero in almost all the prior estimations we made. We had to limit wage wedge split up between internal terms of exchange and fiscal and social wedge without being able to split up within the latter. In other respects, the most satisfactory models were obtained using hourly labour cost and productivity specifications (and not per capita). Finally, modelling attempts with unemployment rate rather than its logarithm were unsuccessful.

[3] See Rault (2000) for a discussion on weak exogeneity and causality.

The model adopted was composed of ten variables (unemployment rate, hourly real cost, hourly productivity, replacement ratio, mismatch, real interest rate, quit ratio, working hours, the terms of exchange, fiscal and social wedge (which combine four deduction rates)). The variable formulation of the statistical model stated by equation (1) is given by the vector X_t = (*u*, *w-p*, *prodh*, *tr*, *mm*, *r*, *ec*, *h*, *pc-p*, *coinfs*)'$_t$. Its purpose is to study the interdependences between these variables, transformed in natural logarithm, without making any *a priori* hypothesis on the value of the elasticities linking them and to test the existence of long run relations.

Two Co-integration Relations

The lag length choice used in the specification of the unrestricted VAR-ECM model is based on the results of two information criteria (Schwarz's Bayesian information criterion and the Hannan-Quinn criterion), and on global Fisher's tests. These different methods all indicate an optimal value of two quarters. One must notice that the lag length choice used in the VAR-ECM model is a crucial stage of the analysis, since it can noticeably affect the determination of the dimension of the co-integrating space, that is, the rank of the Π matrix: simulations by Boswijk and Franses (1992), and Gonzalo (1994) show that under-fitting leads to underestimating the number of long run relations, whereas over-fitting leads to overestimating this number. Moreover, these simulations show that asymptotic distributions of the trace and eigenvalue tests proposed by Johansen (1988), can be rather bad approximations of the true small sample distributions, and should therefore be used with caution. Boswijk and Franses (1992) advocate using the corrected version of these two tests, which perform better in the case of small or medium sample size. These small sample corrected versions of test statistics denoted by $\lambda_{\max}^{adj}$ and λ_{trace}^{adj}, are obtained by pre-multiplying the usual test statistics by (T - np) instead of T, where n is the model variable number and p the VAR order.

Once the lag length used in VAR-ECM model specification has been determined, the next step is to test the number of co-integrating relationships existing between the ten variables of the system. At this stage, one aforementioned point must be emphasised: the asymptotic distributions of the co-integration tests depend on the deterministic components (which are not explicitly modelled) in the system. Specifically, these tests depend on the possible presence of a constant or linear deterministic trend in the long run relations. For instance, if the linear deterministic trend is not constrained to lie in the co-integrating space, the presence of a non-zero deterministic trend outside the long run relations indicates the presence of a quadratic trend in every component of the system taken in level, since the system is written in first differences. In the same way, if the constant is unrestricted, this modelling allows for a linear deterministic trend in the level of series.

To know how to model these deterministic components, one can possibly use the results of the sequences of standard unit root tests applied previously, especially the Schmidt-Phillips (1992) ones, which have not eliminated the possibility that some of these series have a linear drift. That's why all the co-integrating rank tests have been investigated in a system with an unrestricted constant, as well as a linear deterministic trend constrained to lie in the co-integrating space. The small sample corrected versions of the two LR test statistics (trace test and Lambda max test) and also the critical value taken from Johansen (1995), are reported in Table 2.

Table 2. Estimation of the Number of Co-integrating Relationships

Ho against Ha	$\lambda_{\max}^{adj}$		λ_{trace}^{adj}	
	Statistic	Critical value[a]	Statistic	Critical value[a]
r = 0 against r = 1	77.22 **	66.2	310.90 **	263.4
r ≤ 1 against r = 2	60.46	61.3	233.60 *	222.2
r ≤ 2 against r = 3	48.07	55.5	173.20	182.8
r ≤ 3 against r = 4	39.97	49.4	125.10	146.8
r ≤ 4 against r = 5	32.50	44.0	85.14	114.9
r ≤ 5 against r = 6	16.97	37.5	52.64	87.3
r ≤ 6 against r = 7	14.43	31.5	35.66	63.0
r ≤ 7 against r = 8	10.52	25.5	21.23	42.4
r ≤ 8 against r = 9	7.67	19.0	10.71	25.3
r ≤ 9 against r = 10	3.03	12.2	3.037	12.2

Note: [a] critical value at 5 %. ** is significant at 1% level, * is significant at 5% level.

These test statistics indicate the existence of two co-integrating relationships between the ten variables considered.[4 ,5] The estimation of the co-integrating vectors and of the adjustment coefficients will be given later.

Once the co-integrating rank was determined, systematic LR tests on the deterministic components were made. These tests confirmed the results and led to the acceptance of a specification of the Vector Error Correction Model (VAR-ECM), with an unrestricted constant in the short run, as well as a linear deterministic trend constrained to lie in co-integrating relationships. From here on model specification was completely determined (two lags, two co-integrating relationships and a linear deterministic trend constrained to lie in co-integrating relationships).

D. Weakly Exogenous Variables and that Excluded from Co-integrating Space

The next step is to ask oneself if some system variables can be considered as weakly exogenous for the parameters of the two co-integrating relationships found previously. If so, these parameters can then be estimated without loss of information from the more manageable conditional model, having been extracted from the full VAR-ECM model. This hypothesis of weak exogeneity is expressed by the nullity of some coefficients of the α matrix. Table 3 produces the results of these weak exogeneity tests.

The results can be synthesised as follows: at a 5 % level, one rejects the weak exogeneity of real labour cost, of unemployment rate, of working hours, of mismatch, of the terms of exchange, of hourly productivity and of quit ratio. Moreover, at a 5 % level, the joint weak exogeneity hypothesis of the remaining three variables is easily accepted by the data ($\chi^2(6) = 5.24$ (0.51)). Therefore, we chose to estimate the two long run relations from a partial VAR-ECM model composed of seven equations (*w-p*, *u*, *h*, *mm*, *pc-p*, *prodh*, *ec*), conditional to

[4] The outcome of the co-integration analysis remains unchanged if we use the critical values recently tabulated by Pesaran, Shin and Smith (1999).

[5] Given that the calculated statistical value of the $\lambda_{\max}^{adj}$ test is very close to the 5 % critical value, it is reasonable to think as economic theory suggests, that there exist two long run relationships between the considered variables: that is what it indicates in addition to the λ_{trace}^{adj} test.

Table 3. Weak Exogeneity Tests of the Different Variables for all Long Run (α and β) Parameters

Variable	Weak exogeneity	LR test statistic
w - p	rejected	χ^2 (2) = 19.13 (0.00)
u	rejected	χ^2 (2) = 11.39 (0.00)
tr	not rejected	χ^2 (2) = 2.56 (0.27)
r	not rejected	χ^2 (2) = 0.97 (0.61)
coinfs	not rejected	χ^2 (2) = 4.03 (0.13)
h	rejected	χ^2 (2) = 19.27 (0.00)
mm	rejected	χ^2 (2) = 17.23 (0.00)
pc - p	rejected	χ^2 (2) = 12.84 (0.00)
prodh	rejected	χ^2 (2) = 10.78 (0.00)
ec	rejected	χ^2 (2) = 27.98 (0.00)

Note : The number in brackets indicates the marginal asymptotic level, namely the probability of exceeding the value of the computed statistic. Thus a marginal asymptotic level of 27 % (0.27), for instance, means that for an α level smaller than 27 %, the null hypothesis Ho of weak exogeneity of the variable under study is accepted.

the three equations describing the evolution of the weakly exogenous variables (*tr*, *r*, *coinfs*).

Then a first sequence of tests was applied in order to determine if some system variables could be considered excluded from the two long run relations. The following table shows that at a 5% level, replacement rate, real interest rate and working hours do not belong to the co-integrating space. Moreover at a 5 % level, the joint exclusion hypothesis of these three variables of the co-integrating space is easily accepted by data ($\chi^2(6) = 2.30$ (0.89)). The replacement ratio and the real interest rate are thus both weakly exogenous and excluded from the co-integrating space, which in other words means that they only have an influence on the short run dynamic of the price and wage schedule.

Next it is interesting to ask oneself if there exists a variable belonging to the co-integrating space, which constitutes a co-integration relation alone. In this respect, Table 5 presents the results of the stationarity tests around a linear deterministic trend of the different variables. For instance, to test if the

Table 4.Tests of the Structure of Co-integrating Space

Variable	Belonging to co-integrating space	LR test statistic
w - p	yes	χ^2 (2) = 31.46 (0.00)
u	yes	χ^2 (2) = 15.91 (0.00)
tr	no	χ^2 (2) = 0.19 (0.90)
r	no	χ^2 (2) = 1.12 (0.57)
h	no	χ^2 (2) = 0.50 (0.77)
coinfs	yes	χ^2 (2) = 6.36 (0.04)
pc - p	yes	χ^2 (2) = 6.97 (0.03)
prodh	yes	χ^2 (2) = 6.39 (0.04)
ec	yes	χ^2 (2) = 26.15 (0.00)
trend	yes	χ^2 (2) = 6.46 (0.03)

Notes: a) Some of the results given in this table were obtained after several iterations. In fact, two weekly exogenous variables were shown moreover not to belong to the co-integrating space. We found it more logical to take these two pieces of information into account step by step, instead of directly placing these two variables in the short run. For this purpose, we first estimated a VAR-ECM in which the replacement rate only belonged in the short run dynamic, then re-tested in this framework, to see if the other variables belonged to the co-integrating space. b) The number in brackets indicates the marginal asymptotic level, namely the probability of exceeding the value of the computed statistic. Thus a marginal asymptotic level of 90 % (0.90) for instance, means that for an α level smaller than 90 %, the null hypothesis Ho of exclusion from the co-integrating space of the variable under study is accepted by the data.

unemployment rate *u* is stationary around a linear deterministic trend, one has to test if vector b' = (0 1 0 0 0 0 0 a) belongs to the co-integrating space. The results of these tests are categorical, since they reject the stationarity hypothesis around a linear deterministic trend of the seven variables belonging to the co-integrating space in every case. Thus, the results of the stationarity tests applied in the multivariate framework, where the interdependences between variables are explicitly modelled, are concomitant with those applied previously in the univariate framework. These tests indicate that the variables are characterised by a stochastic non-stationarity (namely integrated of order 1), rather than a deterministic non-stationarity (namely stationary around a linear deterministic trend).

Table 5. Stationarity Tests of the Different Variables Around a Linear Deterministic Trend

Variable	Stationarity around a linear deterministic trend	LR test statistic
w - *p*	rejected	χ^2 (6) = 33.11 (0.00)
u	rejected	χ^2 (6) = 31.02 (0.00)
mm	rejected	χ^2 (6) = 52.65 (0.00)
coinfs	rejected	χ^2 (6) = 29.74 (0.00)
pc - *p*	rejected	χ^2 (6) = 58.59 (0.00)
prodh	rejected	χ^2 (6) = 41.84 (0.00)
ec	rejected	χ^2 (6) = 34.03 (0.00)

Table 6 gives the estimation of the two long run relations and the error correction coefficients obtained from the conditional model.

E. PS and WS Identification

Spontaneously, each of the two co-integrating vectors has an unemployment rate coefficient with an opposite sign, which indicates both a price and wage setting behaviour. Nevertheless, it is important to notice that these two co-integrating vectors have no economic meaning at this stage, and are nothing other than a vectorial basis of the co-integrating space. Strictly speaking, they are obtained as the eigenvectors of the long run Π matrix and any linear combination of these two vectors forms a new co-integrating relationship between the seven variables. These vectors then have only a purely statistical value. Econometric modelling alone does not allow the structural form of (WS) and (PS) curves to be determined ex nihilo. Therefore, it does not eliminate a theoretical consideration of the form of structural equations, but requires on the contrary, the a priori specification of identification conditions,

Table 6. Maximum Likelihood Estimations of the Normalised Co-integrating Vectors and of the Error Correction Coefficients

Variables	Normalised co-integrating vectors (β matrix)	
w - p	1.000	1.000
u	0.254	-0.506
mm	-0.083	-0.000
pc - p	-0.733	1.042
prodh	0.087	-3.012
ec	-0.403	0.260
coinfs	0.764	1.642
trend	-0.001	0.014
Variables	**Error correction coefficients (α matrix)**	
w - p	-0.091	0.087
	(-3.84)	(6.77)
u	0.047	0.155
	(1.73)	(4.50)
mm	0.294	0.054
	(3.52)	(1.20)
h	-0.062	-0.034
	(-3.52)	(-4.06)
pc - p	-0.045	0.053
	(-1.64)	(3.48)
prodh	-0.042	0.068
	(-1.96)	(4.06)
ec	0.430	0.122
	(5.10)	(2.40)

Note: The number in brackets represents the t stats.

using a theoretical model, before beginning the estimation. The identification of the two curves is investigated here using the following two theoretical restrictions: the wage determination (WS curve) is supposed to be made independently of productivity level (the Manning, 1993, identification restriction) and unemployment is not supposed to influence wage determination (PS curve). Structural forms are then obtained by calculating the two linear combinations of the estimated co-integrating vectors, which satisfy identification constraints. It must be emphasised that it is not a test, but simply a change of basis in the co-integrating space, in order to statistically distinguish between the two structural equations. Thus, these constraints do not affect the level and evolution of equilibrium unemployment estimation (which is robust to identification choice). After normalisation, the two (just) identified long run relations are given by,

$$\begin{aligned}
\text{(PS)} \quad w - p &= 0.055\, mm + 0.138\, pc\text{-}p + 0.944\, prodh - 0.041\, coinfs \qquad (2)\\
&\quad + 0.181\, ec - 0.004\, trend\\
\text{(WS)} \quad w - p &= -0.232\, u + 0.080\, mm + 0.679\, pc\text{-}p + 0.693\, coinfs\\
&\quad + 0.384\, ec - 0.001\, trend
\end{aligned}$$

Finally, over-identifying restrictions were tested, the results are reported in Table 7: the exclusion of the fiscal and social wedge, of the terms of exchange and of the linear deterministic trend from the PS curve are accepted at a 5% level.

Additional structural hypotheses were also tested, as the exclusion of *mm* and *ec* variables from (PS), but were all rejected. The presence of these variables in price equation is not theoretically justified, which is one reason for dissatisfaction. Finally, the two over-identified long run relations are given by:

$$\begin{aligned}
\text{(PS)} \quad w - p &= 0.073\, mm + 0.204\, prodh + 0.230\, ec \qquad (3)\\
\text{(WS)} \quad w - p &= -0.050\, u + 0.078\, mm + 0.117\, pc\text{-}p + 0.159\, coinfs\\
&\quad + 0.274\, ec + 0.001\, trend
\end{aligned}$$

Table 7. Tests of Over-identifying Restrictions

Null hypothesis	Accepted hypothesis	LR test statistic
Exclusion of h from (PS) and (WS), and exclusion of *pc* - *p* from (PS)	yes	χ^2 (3) = 0.94 (0.82)
Exclusion of h from (PS) and (WS), and exclusion of *pc* - *p* and *coinfs* from (PS)	yes	χ^2 (4) = 0.95 (0.92)
Exclusion of h from (PS) and (WS), and exclusion of *pc* - *p*, *coinfs* and of the linear deterministic trend from (PS)	yes	χ^2 (5) = 6.21 (0.29)

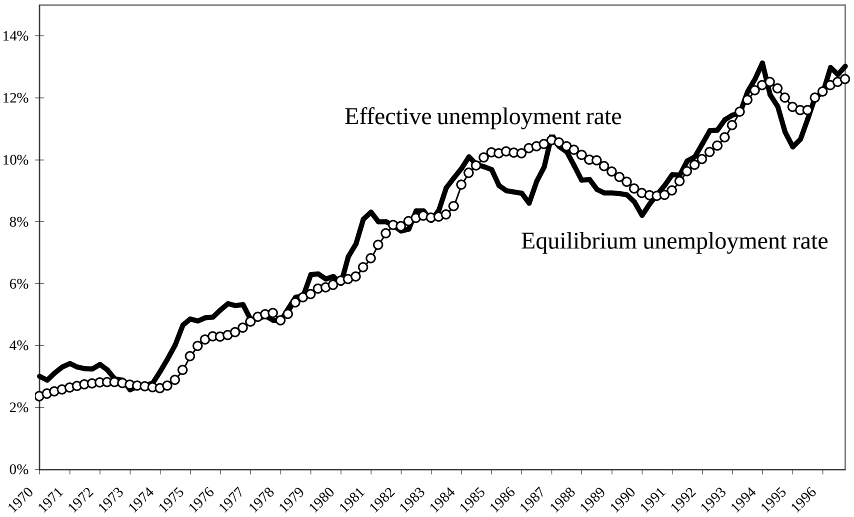
It is now possible to determine the equilibrium unemployment from the two estimated structural equations. For this purpose, one must resolve the equilibrium of the partial system of the labour market obtained. This resolution gives the following expression of equilibrium unemployment.

$$\begin{aligned} u^* \; = \; & -4.1\, prodh + 2.34\, pc - p + 0.1\, mm + 0.88\, ec \\ & + 3.18\, coinfs + 0.02\, trend \end{aligned} \tag{4}$$

All equilibrium unemployment determinants have a sign in accordance with the theoretical idea. Equilibrium unemployment decreases when productivity growth exceeds the trend, which corresponds to an annual growth rate of over 2% (this is close to the average rate of productivity growth over the period covered). Unemployment increases with the terms of exchange (the oil crisis for instance has increased unemployment, since it led to a higher rise in consumption prices than added value prices), with the growth of skill mismatch, quit ratio, fiscal and social wedge and its components. The contributions of the terms of exchange and of mismatch remain quite small (about 5% of the equilibrium unemployment increase).

Figure 1 represents effective unemployment rate and equilibrium unemployment rate. The latter is defined up to a constant, which requires a choice in reference value: we choose the 1973 average rate, so we assumed equality between effective unemployment and equilibrium unemployment in that year. Neither equilibrium unemployment nor its determinants were smoothed here.

Figure 1. Effective Unemployment Rate and Equilibrium Unemployment Rate



F. Diagnostic Tests on the Residuals

The last step is to establish whether the estimated VAR-ECM model is a reasonably congruent representation of the data. We have therefore implemented two kinds of tests: misspecification and constancy tests.

Firstly, several test statistics were calculated in order to check the quality of the multivariate estimation (Lagrange Multiplicator (LM) test and Ljung-Box test for serial correlation of order 16, ARCH (Autoregressive Conditional Heteroscedasticity) tests, Jarque-Bera normality test). The tests constitute a good way to detect possible failings of some hypotheses made during the system estimation. These tests indicate that the conditional VAR-ECM model is well behaved and not subject to misspecification, since the usual hypotheses concerning the residuals of each of the seven equations are verified (see Table 8).[6]

Table 8. Specification Tests of the Residuals of the Conditional VAR Model

Equation	LB (16)	WHITE (F-Form)	ARCH (16)	JB (2)
Dw - p	19.43	0.69	20.15	1.59
	(0.14)	(0.87)	(0.21)	(0.44)
Du	14.64	1.37	18.99	32.21
	(0.40)	(0.16)	(0.26)	(0.00)
Dmm	17.03	1.57	15.81	4.53
	(0.25)	(0.07)	(0.46)	(0.10)
Dh	24.55	0.67	24.85	61.39
	(0.03)	(0.93)	(0.07)	(0.00)
Dpc - p	30.23	0.98	23.79	4.21
	(0.007)	(0.52)	(0.09)	(0.12)
Dprodh	11.69	0.56	11.74	5.68
	(0.63)	(0.92)	(0.76)	(0.05)
Dec	21.87	1.01	13.86	75.01
	(0.08)	(0.48)	(0.60)	(0.00)

Note: The number in brackets indicates the marginal asymptotic level, namely the probability to exceed the value of the computed statistic. Thus a marginal asymptotic level of 14 % (0.14) for instance, means that for a Ho level smaller than 14 %, the null hypothesis Ho of absence of residual serial correlation of order 16 is accepted by data.

[6] The residuals of the conditional VAR-ECM model equations have good properties on the whole: they do not suffer from serial correlation, are not of ARCH type, even if they sometimes have normality problems. This lack of normality assumption in some equations is not actually very serious for the conclusions of the study, since as noted by Johansen (1995), the asymptotic properties of the Maximum Likelihood method only depend on the i.i.d assumption of the errors.

Secondly, the conditional and marginal VAR-ECM models were re-estimated by recursive least squares until 1996/4 and One Step Ahead, as well as performing Backward and Forward Chow tests in order to appreciate the parameter constancy through time. The graph examination does not reveal any particular break and was not reported here.

Thus, the misspecification and constancy tests indicate the estimated conditional VAR-ECM model to be a satisfactory representation of the data.

IV. Conclusion

One can consider a great number of possible explanations as to the rise and persistency of unemployment in France. The aim of this paper was to confront some of these determinants with data in a WS-PS model estimation framework on French macroeconomic data.

First and foremost, we chose a selection of about fifteen variables whose influence rested on explicit micro-economic bases and which was founded on a general equilibrium framework. To this first filter, of a theoretical order, a second one of a statistical order was added, resulting in the possibility of building indicators for these determinants, then a third one of an econometric order was added, resulting in the model estimation. Finally, only five variables reached the end of this procedure. The equilibrium unemployment increase in France reflects the slowing down of productivity gains, the increase of social and fiscal wedges, the deterioration in job security and in a more marginal way, the terms of exchange increase and the skill mismatch.

Considering a richer set of variables and a different methodology, this paper confirms the impact of some unemployment determinants in a unified framework, found in previous studies incorporating a limited number of candidates to explain equilibrium unemployment (Bonnet and Mahfouz, 1996; L'Horty and Sobczak, 1997; Cotis, Méary and Sobczak, 1997). It gives a main role to the rise of social and fiscal wedge, as do two of the previous studies (LS, 1997 and CMS, 1997). It is also compatible with a predominant role attributed to the influence of real interest rate, when this influence is well mediated by a downturn in productivity gain, also in keeping with the three studies. It also concludes that the terms of exchange play a role in the formation of French unemployment, like one of the studies (BM, 1996). Our empirical investigation also shows the influence of skill mismatch and of the employment protection system, via the quit ratio, which has not been obtained (nor introduced) before, in the existing applied French studies using time series. In addition, our study leads to a questioning of the influence of numerous other determinants: the replacement rate would not have had any impact on the increase of equilibrium unemployment (contrary to the LS, 1997, results), and would be the same for other determinants which were not introduced in previous studies: the lesser digressiveness of social wedge, the reduction of working hours and the minimum wage increase.

References

Banerjee, A., Dolado, J., Galbraith, J.W, and D.F Hendry (1993), *Co-integration, Error Correction, and the Econometric Analysis on Non-stationary Data,* Oxford, Oxford University Press.

Bean, C. (1994), "European Unemployment: A Survey," *Journal of Economic Literature* **32**: 573-620.

Binmore, K.G., Rubinstein, A., and A. Wolinsky (1986), "The Nash Solution in Economic Modelling," *Rand Journal of Economics* **17**: 176-88.

Bonnet, X., and S. Mahfouz (1996), "The Influence of Different Specification of the Wage-price Spiral on the Measure of the NAIRU: The Case of France," *Document de Travail 9611*, INSEE.

Boswijk, H.P., and P.H Franses (1992), "Dynamic Specification and Co-integration," *Oxford Bulletin of Economics and Statistics* **54**: 369-381.

Cahuc, P., and A. Zylberberg (1996), *Economie du Travail*, De Boeck Université.

Cahuc, P., and A. Zylberberg (1999), "Le Modèle WS-PS," *Annales d'Economie et de Statistique* **53**: 1-30.

Cotis, J.-Ph., Méary, R., and N. Sobczak. (1998), "Le Chômage d'Equilibre en France: Une Evaluation," *Revue Economique* **49**: 921-935.

Friedman, M. (1968), "The Role of Monetary Policy," *American Economic Review* **58**: 1-17.

Heyer, E., Le Bihan, H., and F. Lerais (2000), "Relation de Phillips, Boucle Prix-salaire: Une Estimation par la Méthode de Johansen," *Economie et Prévision* **146**: 43-60.

Jackman, R., Layard, R., and S. Savouri (1991), "Mismatch: a Framework for Thought," in F. Padoa Schioppa, ed., *Mismatch and Labour Mobility*, Cambridge, Cambridge University Press.

Jakobsson, U. (1976), "On the Measure of the Degree of Progression," *Journal of Public Economics* **5**: 161-68.

Jobert, T. (1992), "Test de Racine Unitaire: Une Stratégie et sa Mise en Oeuvre,*"* *Cahiers Eco & Maths* **92-44**, Université Paris I.

Johansen, S. (1988), "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control* **12**: 231-254.

Johansen, S. (1995)*, Likelihood-based Inference in Co-integrated Vector Autoregressive Models*, Oxford University Press.

Johansen, S., and K. Juselius (1994), "Identification of the Long-run and the Short-run Structure: An Application to the IS-LM Model," *Journal of Econometrics* **63**: 7-36.

Koskela, E., and J. Vilmunen (1994), "Tax Progression is Good for Employment in Popular Models of Trade Union Behaviour," *Discussion Papers 3/94*, Bank of Finland.

Kwiatkowski, D., Phillips, P.C.B, and Y. Shin (1992), "Testing for the Null Hypothesis of Stationarity against the Alternative of a Unit Root," *Journal of Econometrics* **54**: 159-178.

L'Horty, Y., and N. Sobczak (1997), "Les Déterminants du Chômage d'Equilibre: Estimation d'un Modèle WS-PS," *Economie et Prévision* **127**: 101-16.

Laffargue, J.P. (1995), "A Dynamic Model of The French Economy, with Rational Expectations, Monopolistic Competition, and Labour Market Bargaining," *Annales d'Economie et de Statistique* **37-38**: 465-530.

Layard, R., Nickell, S., and R. Jackman (1991), *Unemployment: Macroeconomic Performance and the Labour Market*, Oxford, Oxford University Press.

Layard, R, and S. Nickell (2000), "Labor Market Institutions and Economic Performance," in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics 3C*, New York and Oxford, Elsevier Science, North-Holland,.

Lindbeck, A. (1993), *Unemployment and Macroeconomics*, Cambridge, MA, MIT Press.

Lockwood, B., and A. Manning (1993), "Wage Setting and the Tax System:

Theory and the Evidence for United Kingdom," *Journal of Public Economics* **52**: 1-29.

Manning, A. (1993), "Wage Bargaining and the Phillips Curve: The Identification and Specification of Aggregate Wage Equations," *The Economic Journal* **103**: 98-118.

Oswald, A.J. (1985), "The Economic Theory of Trade Unions: An Introductory Survey," *Scandinavian Journal of Economics* **87**: 160-193.

Pesaran, M.H., Shin, Y., and J.S. Smith (1999), "Structural Analysis of Vector Error Correction Models with Exogenous I(1) Variables," *manuscript*, University of Cambridge.

Phelps, E. (1994), *Structural Slumps, The Modern Equilibrium Theory of Unemployment, Interest, and Assets*, Harvard University Press.

Pissarides, C.A. (1990), *Equilibrium Unemployment Theory*, Oxford, Basis Blackwell.

Rault, C. (1997), "Prédétermination, Causalité, Exogénéité dans un Modèle Vectoriel à Correction d'Erreur: Identifiabilité d'une Forme Structurelle," *Cahiers Eco & Maths* **97-60**, Université Paris I.

Rault, C. (2000), "Non-causality in VAR-ECM Models with Purely Exogenous Long Run Paths," *Economics Letters* **67**: 121-129.

Schmidt, P., and Phillips, P.C.B (1992), "LM Tests for a Unit Root in the Presence of Determinist Trends," *Oxford Bulletin of Economics and Statistics* **54**: 257-287.

Sneessens, H.R. (1994), "Courbe de Beveridge et Demande de Qualifications," *Economie et Prévision* **113-114**: 127-138.

Unedic (1997), "Quelle Mesure du Taux de Remplacement?," *Bulletin de Liaison* **145**: 152-157.

Journal of Applied Economics, Vol. VI, No. 1 (May 2003), 157-175

HOW CAN WE USE THE RESULT FROM A DEA ANALYSIS? IDENTIFICATION OF FIRM-RELEVANT REFERENCE UNITS

Jonas Månsson[*]
Växjö University

Submitted August 2002; accepted February 2003

Two types of guidelines can be obtained from a DEA (data envelopment analysis) analysis. Firstly, the firm can reduce input or increase production according to the DEA results. Secondly, an inefficient firm might be able to identify reference units. This makes it possible for the inefficient firm to, on site, study production that is more efficient, and thereby get information on e.g. efficient organisational solutions. In this study, we focus on how to detect these firm-relevant reference units. While applying the existing methods for identification of reference units, i.e. the *intensity variable* method and the *dominance method*, on a data set concerning booking centres in the Swedish taxi market, shortcomings in these methods were identified. This motivates the development of a new method. This new method, the *sphere measure*, enables an inefficient unit to identify existing and efficient units that have the largest similarity with itself. The identified units will thus be firm-relevant reference units.

JEL classification codes: D24, L25
Key words: reference units, firm-relevant, DEA

I. Introduction

There are two kinds of guidelines that can be provided to firms as a result of a DEA-analysis on technical efficiency.[1] First, one guideline would be

[*] I would like to thank Prof. Rolf Färe, Oregon State University, Prof. Lars Gunnar Svensson, Lund University and Dr. Ann Veiderpass, Gothenburg University for valuable comments on previous versions of this paper. Correspondence should be addressed to: Department of Economics and Statistics, Växjö School of Management & Economics, Växjö University, S-351 95 Växjö, Sweden.

[1] In data envelopment analysis, DEA (see e.g. Charnes, Cooper and Rhodes, 1978), the

how much a specific unit will be able to reduce its input while still being able to produce the same amount of output. This type of guideline does not take technical or organisational obstacles into consideration.[2] Therefore, a second type of guideline is to identify units that can serve as a reference for an inefficient unit.[3] Relevant reference units make it possible for inefficient units to, on site, study production that is more efficient than its own. This makes it possible to adopt more efficient ways to organise production.

In the literature, two methods are discussed as a means to identify reference units based on the result of a DEA analysis. These are the *intensity variable method* (See Kittelsen and Førsund, 1992) and the *dominance method* (See Tulkens, 1993). We have explored these two methods on a data set concerning the production of booking centre services in Sweden, and identified shortcomings in these methods. In some cases, units, which were defined as reference units for a specific inefficient unit, had little similarity with regard to amount of input used and output produced. Results of this type that are reported to managers will undermine confidence in the DEA method. Furthermore, while investigating the dominance method another shortcoming was identified. For some units, it was not possible to identify a reference unit that dominated the inefficient unit. The identified shortcomings in the existing methods of detecting reference units, for an inefficient unit, motivate the search for a new method. The starting point for this search is to list properties that are desired for reference units. Then we use these properties to construct a measure/method that fulfils these properties.

reference technology, is specified as an activity analysis model (see e.g. von Neumann, 1938). The model is also referred to as the non-parametric method (see e.g. Färe, Grosskopf and Lovell, 1985). The input based framework used in this study originates from Farrell (1957) and was later generalised to also cover non-homogeneous production technologies, i.e., allowing for variable returns to scale, by Førsund and Hjalmarsson (1974,1979). The idea was presented in 1974 and implemented in 1979. In Färe, Grosskopf and Lovell (1983), the framework was further generalised to cover multiple output and different disposability assumptions.

[2] For example, a small unit may find it efficient to handle administrative issues manually, while large units computerise.

[3] This is unlikely to happen in a competitive environment, but in e.g. the public sector, providing this information to others may not be a problem.

The outline of this study is as follows. In Section II, we will state the framework used in the study. We start by set up the DEA problem and presenting a list of desired properties. These properties are as follows. A reference unit should always exist, the reference unit should be efficient, the reference unit should be an existing unit (i.e. excluding hypothetical reference units such as convex combinations of existing units), and finally a reference unit should be as similar as possible to the inefficient unit. Data is presented in Section III. In Section IV, we first evaluate the existing methods with respect to the desired properties presented in Section II. As mentioned above, we could show that in some cases, designated reference units had little similarity with the inefficient unit. In the case of the dominance method, we could also show that reference units in some cases did not exist. We therefore introduce a new method, the s*phere measure*, which is constructed so that it will fulfil the desired properties. The method will guarantee the existence of a unit, chosen among existing efficient units so that it will minimise the Euclidean norm between the reference unit and the inefficient unit, i.e. has the largest similarity. In Section V, the results are summarised and some concluding remarks are stated.

II. Framework

A. Measuring Efficiency with DEA

Since the aim of this study is to state desired properties of a reference unit, as a result of a DEA analysis, we first need to set up the DEA problem. Let there be $k = 1,\ldots, K$ observations, x_{kn} inputs $n = 1,\ldots, N$, and y_{km} outputs, $m = 1,\ldots,M$. The vector to be enveloped for observation k is then $(x_k, y_k) = (x_{k1},\ldots, x_{kN}, y_{k1},\ldots, y_{kM})$. Then the programming problem to be solved for a unit k' is as follows

$$TE(x_{k'}, y_{k'}) = Min\ \lambda_{k'} \qquad (1)$$

s.t.

$$i) \qquad \sum_{k=1}^{K} z_k y_{km} \geqq y_{k'm},\ m = 1,\ldots,M$$

ii) $$\sum_{k=1}^{K} z_k x_{kn} \leqq \lambda_{k'} x_{k'n},\ n = 1,...,N$$

iii) $$\sum_{k=1}^{K} z_k = 1$$

iv) $$z_k \geqq 0,\ k = 1,...,K$$

where $\lambda_{k'}$ is the efficiency score to be calculated. Since an input based framework is used, the minimum of $\lambda_{k'}$ equals the largest possible contraction of the input vector, such that the unit still remains in the reference technology. We also assume strong disposability of both inputs and outputs and a variable return to scale technology. The latter is given by restriction *iii*.

B. Desired Properties of Reference Units

Before stating and discussing desired properties of a reference unit, some definitions and notations have to be made. First, denote the set of all observed units by **K** = {1,…, *k, K*}. The set of reference units for a specific unit *k* is denoted $\Re e_k$, i.e. if unit *j* is a reference unit for unit *k*, then $j \in \Re e_k$. Finally, given an input requirement set *L(y)*, we can define the isoquant of this input requirement set as *Isoq L(y)* = {*x* : *x* ∈ *L(y)*, λ *x* ∉ *L(y)* for all $\lambda \in$ [0,1]}. Given the definitions and notations above, we will state desired properties and subsequently discuss them.

Table 1. Desired Properties of a Reference Unit/s

	Property
1	$\Re e_k \neq \varnothing$
2	If unit $j \in \Re e_k$ then $x_j \in Isoq\ L(y)$
3	If unit $j \in \Re e_k$ then unit $j \in \mathbf{K}$
4	If unit $j \in \Re e_k$ then there cannot exist another unit i, $x_i \in Isoq\ L(y)$, such that $\lVert ik \rVert < \lVert jk \rVert$

A first property is that at least one possible reference unit should exist, i.e. $\Re e_k \neq \varnothing$. This property might seem redundant, but as will be discussed later, one of the existing methods may produce results where a reference unit does not exist.

Since a goal for all economic activity is the efficient use of resources, the second property we claim for a reference unit is that it should be efficient. This is given by the second property that states: if unit *j* is a reference for an inefficient unit *k*, i.e., $j \in \Re e_k$ then it is impossible to contract the input vector of unit *j*, while still being able to produce the same amount of outputs, i.e. $x_j \in Isoq\, L(y)$.

Further, the aim of using a reference unit is that it should be possible for an inefficient unit to study the production of the reference unit on site. The third property states that if unit *j* is to be a reference unit for an inefficient unit *k*, i.e. $j \in \Re e_k$, unit *j* has to be observable, i.e. $j \in \mathbf{K}$. Thus, property 2 excludes convex combinations of existing units.

So far we have excluded all other inefficient units and convex combinations of existing efficient units from the possible reference set. However, we are still left with a considerable amount of possible units. From a practical point of view, to make an impact on firms trying to become more efficient, we need to guide them to reference units that in some sense are similar to their own firm. The term similarity is not easy to define since two units can be similar/dissimilar in many different dimensions.[4] However, since DEA analysis is an analysis of production and researchers are likely to at least have information about production data, we therefore define similarity as producing a similar amount of outputs and use a similar amount of inputs. To define similarity in a multidimensional framework, we need a measure that is able to take multidimensionality into consideration. The Euclidean norm is one such measure and will here be used as a measure of similarity. Further, we will claim that the most similar unit among possible reference units is most suitable reference unit. Thus, the forth desired property of a reference unit *j* is that another possible reference unit $i \in \Re e_k$ there should not exist, such that the distance between unit *i* and the inefficient unit *k* is smaller than the distance between unit *j* and unit *k*, i.e. $\|jk\| < \|ik\|$ for all *i*.

[4] E.g., two units can be similar with respect to location, education of management, gender representation etc.

Given the properties above, we now turn to empirically explore these properties related to a data set. We start by exploring the two existing methods, the *intensity variable method* and the *dominance method*, and finally we introduce a new method labelled the *sphere measure*.

III. Data

The data in this study concerns production of booking centre services in the Swedish taxi market. The data was collected and confirmed on site at the booking centres during a three-week period in March 1994 and later used in Althin, Färe and Månsson (1994).[5] The production of booking centre services consists of two outputs. The first output is a measure of directly mediated service (*Y1*), i.e. a person orders a taxi and the booking centre immediately mediates the order to a taxi vehicle. The production of the second output, number of co-ordinated and mediated services (*Y2*), is carried out in two steps. The first step is that a person orders a taxi. The order will be co-ordinated with other orders, either by placing more than one customer in the taxi vehicle or by re-directing the taxi vehicle to minimise non yielding transportation. After co-ordination, the order is mediated to the taxi vehicle.

The inputs are:

X1: Number of hours worked annually by personnel directly involved with booking and mediation.
X2: Numbers of hours worked annually by administrative staff.
X3: Number of telephone lines to the booking centre. This will serve as a measure of technical capacity.
X4: Square meters of floor space used for booking services.
X5: Square meters of floor space used for administration.
X6: Value of purchased services in Swedish kronor (SEK).

Descriptive statistics on input and output are presented in Table 2.

A few comments have to be made concerning the data. One can see that there are booking centres that only produce one of the outputs. This can be explained by the fact that the data covers both privately owned and publicly owned

[5] For a more extensive discussion on booking centre production, see Månsson (1996).

Table 2. Descriptive Statistics on Inputs and Outputs for the Production of Booking Centre Services (N = 30)

	Mean	Std. Dev.	Min	Max
Output				
Directly mediated services (Y1)	176,720	208,677	0	1,000,000
Co-ordinated and mediated services (Y2)	77,701	99,041	0	400,000
Input				
Hours worked with booking - mediation (X1)	11,226	10,302	979	54,136
Hours worked with administration (X2)	3,648	4,410	0	20,976
Telephone lines to the booking centre (X3)	10	7.12	1	28
Floor space used for the booking services (X4)	35	35.7	6	200
Floor space used for administration (X5)	40	56.7	0	300
Value of purchased services in SEK (X6)	99,000	271,753	0	1500,000

booking centres. One of the objectives with introducing publicly owned booking centres was to increase the number of co-ordinated services. This explains why *Y1* for some booking centres is zero. On the other hand, the most likely way to administrate an order during the period when the Swedish taxi market was regulated was to mediate the order at the same moment a customer placed the order in the booking centre. Some privately owned booking centres still apply this system, and thereby do not allocate resources to co-ordinate services. This explains the zero value for *Y2.* Zero input values can partly be explained by the fact that some booking centres do not have any administrative staff, instead they buy administrative services. This is most likely to happen in the case of small booking centres.

IV. Empirical Investigation

We first present here the computed efficiency scores. Thereby all units that fulfil property 2 and property 3, i.e., all existing efficient units, are identified. We then apply the existing methods, *dominance* and *intensity variable method*, on the data presented in Section III. As will be seen, both existing methods have some shortcomings as regards desired properties. We therefore propose a new method, which will be labelled the *sphere measure*.

A. Identification of Existing and Efficient Units

The framework presented in Section II was used to compute the efficiency scores. The results of these computations are presented in Table 3.

Table 3. Technical Efficiency, Variable Returns to Scale

Unit no.	Efficiency score	Unit no.	Efficiency score	Unit no.	Efficiency score
1	1.000	11	1.000	21	1.000
2	0.875	12	0.980	22	0.663
3	0.722	13	0.523	23	0.490
4	1.000	14	0.748	24	1.000
5	1.000	15	0.769	25	0.797
6	1.000	16	1.000	26	1.000
7	1.000	17	1.000	27	0.793
8	0.584	18	0.901	28	0.694
9	0.806	19	0.950	29	1.000
10	0.641	20	1.000	30	0.758

As seen in the Table, thirteen units are efficient. The minimum efficiency is 0.49 for unit number 23. This means that unit 23 would have to decrease its inputs by 51 percent in order to become efficient. The mean efficiency score is 0.86, i.e. 14 percent inefficiency, and the standard deviation is 0.16. All units that are technically efficient, i.e. have an efficiency score equal to one, fulfil property 2 and are thus potential reference units. Further, they also fulfil property 3, i.e. are existing units.

B. Existing Methods for Detecting Reference Units

B.1. Intensity Variables

When the non-parametric method is used to compute technical efficiency, inefficient units are compared to a convex combination of efficient units. By investigating the value of the individual intensity variables (z_k), obtained when solving the efficiency problem presented in equation (1), it is possible to identify those units that are used in the construction of the efficiency frontier. According to Kittelsen and Førsund (1992), p.302, this information can be used to select a reference unit among the efficient units.[6]

In Table 4 below, the values of the non-zero intensity variables are presented for the inefficient units. These results can be used to provide the inefficient unit information on which efficient unit it is compared to. For example, the inefficient unit 9 is compared to efficient units 1, 7, 11 and 29. According to the values of the intensity variable, efficient unit 11 is the most relevant reference unit, since it has the highest value on the intensity variable (0.754).

One problem with this method occurs when the most influential unit has very little similarity with the inefficient unit.[7] One way of handling this drawback would be to report all units with non-zero intensity variables. It does not solve the problem, but it will provide the inefficient units with alternative units to be compared with. Another way is to determine some criteria for similarity and investigate if the designated unit is the most similar reference unit.

[6] When using the approach suggested by Kittelsen and Førsund, it is possible that more than one reference unit exists. This will be the case if two, or more units have the same value on their intensity variables.

[7] As can be seen in the Appendix, unit 11 is using much less input and produces much less output in each dimension. My experience is that reporting this type of information back to managers will induce suspicion and undermine creditability of the method, since managers will not see unit 11 as a relevant reference unit.

Table 4. Inefficient Units, Units Used in the Reference Frontier for the Inefficient Unit (Frontier Unit), and the Values of the Intensity Variable

Inefficient unit no.	Frontier unit. no	Intensity variable	Inefficient unit no.	Frontier unit no.	Intensity variable	Inefficient unit no.	Frontier unit no.	Intensity variable
2	7	0.2645		20	0.2246		29	0.0588
	11	0.0990		21	0.1182	23	5	0.0519
	24	0.5247		29	0.4906		11	0.3453
	29	0.1118	13	4	0.5563		16	0.4375
3	16	0.6941		11	0.2209		24	0.0255
	24	0.0144		16	0.1943		29	0.1399
	29	0.2915		29	0.0284	25	7	0.0261
8	7	0.1815	14	4	0.1060		11	0.3035
	21	0.5108		16	0.8940		16	0.4711
	29	0.3078	15	7	0.2194		21	0.1697
9	1	0.0779		16	0.6719		29	0.0297
	7	0.0069		29	0.1087	27	5	0.8044
	11	0.7540	18	16	0.5201		24	0.1956
	29	0.1612		24	0.3517	28	6	0.2496
10	1	0.1005		29	0.1282		7	0.5638
	7	0.0174	19	4	0.0146		29	0.1867
	11	0.2852		16	0.8025	30	11	0.7108
	21	0.4904		29	0.1829		16	0.2155
	29	0.1064	22	4	0.3339		24	0.0193
12	11	0.1666		16	0.6073		29	0.0543

Note: As can be noted, the efficient units 17 and 26 are not used as reference for any inefficient unit. The most likely explanation for this is that both these units are unique, in the sense that they are only compared with each other. They are located on either the vertical or the horizontal line segment in Figure 1.

In this study, we use the difference in the Euclidean norm to *indicate* similarity. The Euclidean norm measures the distance between units. The norm between unit *i* and *k* is here defined as:

$$\| \ ik \ \| = \sqrt{\sum_{n=1}^{N} (\frac{x_{ni}}{\overline{x}_n} - \frac{x_{nk}}{\overline{x}_n})^2 + \sum_{m=1}^{M} (\frac{y_{mi}}{\overline{y}_m} - \frac{y_{mk}}{\overline{y}_m})^2} \tag{2}$$

where $\overline{x}_n$ and $\overline{y}_m$ is the mean of n / m.[8]

The criteria we use is that if the norm between unit *j* and unit *k* is smaller than the norm between another unit *i* and unit *k*, i.e. $\|ik\| > \|jk\|$, then unit *j* is more similar to *k* than unit *i* is to unit *k*, and thereby also a more relevant reference unit. We have computed the Euclidean distance between unit 11 and all other observed efficient units and that result is presented in Table 5.

Table 5. Euclidean Distance between Unit No. 11 and all Other Observed and Efficient Units

Efficient unit	Unit. No. 9	Efficient unit	Unit. No. 9
1	2.32	17	6.67
4	2.45	20	2.60
5	2.30	21	2.69
6	2.00	24	2.75
7	1.98	26	2.44
11	2.60	29	6.62
16	2.47		

As shown in the Table there is a unit that have larger similarity to unit 9 than the by intensity variable method detected unit 11.[9] We can thus conclude that that the intensity variable does not fulfil the desired property 4.

[8] The data is normalised since the norm otherwise will be dependent on how the data is measured.

[9] The difference in each input and output dimension, between unit No. 9 and unit No. 7 is reported in the Appendix.

B.2. Dominance

There is one major critique of the non-parametric, or DEA framework presented above. When computing the efficiency score, the inefficient units are compared with convex combinations of efficient units, instead of existing units. As a consequence of this, Tulkens (1993) presented the idea of dominance, which in turn has its roots in Pareto efficiency.[10] In a multiple input, multiple output framework dominance can be defined either from the input, or the output side. Following Tulkens (1993), input dominance is defined as:

Definition: A unit *k* input dominates k', if and only if

$$y_{km} \geqq y_{k'm} \ , \ m = 1,\ldots,M \ \ and \ \ x_{kn} \leq x_{k'n} \ , \ n = 1,\ldots,N$$

That is, unit *k* input dominates *k'*, if unit *k* produces more or equal amount of output compared to *k'* $(\geqq)$ and uses less input in *at least* one dimension $(\leq)$. An alternative version of dominance is strict dominance, taking both inputs and outputs into consideration at the same time.

Definition: A unit *k* strictly dominates unit *k'*, if and only if

$$y_{km} > y_{k'm} \ , \ m = 1,\ldots,M \ \ and \ \ x_{kn} < x_{k'n} \ , \ n = 1,\ldots,N$$

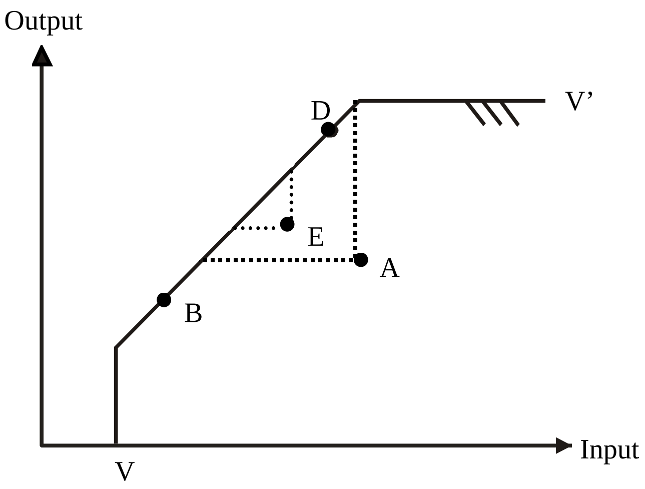
That is, unit *k* strictly dominates unit *k'*, if unit *k* produces more output and uses less input in all dimensions. This means that if unit *k* strictly dominates unit *k'*, then unit *k* also input and output dominates unit *k'*.

As noted by Tulkens (1993), p.191, identification of a dominant unit gives the efficiency score credibility, since it identifies an observed reference unit, instead of a convex combination of existing units.[11] Dominance and a problem with the method are illustrated in Figure 1.

[10] In Tulkens (1993), the author uses the idea of dominance to construct a new reference technology, labelled Free Disposal Hull reference technology (FDH). It should be noted that in this study, we apply the ideas of dominance, given the convexity assumptions of the reference technology, i.e. we do not use the FDH reference technology.

[11] Output dominance is defined analogously, with strong inequality in at least one output dimension.

Figure 1. Illustration of Dominance



In this Figure, unit *A* and unit *E* are inefficient. It is clear that unit *A* is strictly dominated by the efficient unit *D*, since unit *D* uses less input and produces more output than unit *A*. A problem arises, if a situation illustrated by the inefficient unit *E* occurs. Even though unit *E* is inefficient, it is neither dominated by unit *B* nor *D*.[12] Unit *E* produces less output, but at the same time uses less input, compared to unit *D*. The opposite is true when comparing with unit B. Thus, this method may result in a situation where the dominant subset is empty. Dominating references were found for two units for the data used in this study. The efficient unit No. 7 dominated both the inefficient units No. 3 and No. 10. For all other inefficient units, the dominant sub-set was empty, i.e. $\Re e_k \neq \varnothing$. This result was not unexpected, since the model on which the computations were based has as many as 8 dimensions: 2 output dimensions and 6 input dimensions. The more dimensions used in the model, the less likely it is that the dominant subset is non-empty. Thus, the *dominance method* might not fulfil property 1 or property 4.

[12] If the FDH reference technology was used, point E had been considered efficient.

C. The Sphere Measure

In Section IV.B we have demonstrated the intensity variable method and the dominance method and we have identified shortcomings in both methods. We therefore propose a new method with the objective of identifying firm-relevant reference units that fulfil the desired properties listed in Section II.

The idea of the *sphere measure* is rather straightforward. For an inefficient unit, a sphere with radius *r* is defined. The radius of the sphere is then extended until the sphere covers the inefficient unit and at least one efficient unit. The unit that first appears in the interior of the sphere is considered to be the reference unit for the inefficient unit.[13] Moreover, the length of the radius is a measure of how close the inefficient unit is located to the reference unit.

First, denote the subset of efficient observations $\mathbf{S} \subseteq \mathbf{K}$. The subset **S** contains all efficient units from the set of all units, **K**. For an inefficient unit *k,* and an efficient unit $s \in \mathbf{S}$, the radius of the sphere is defined and computed as:

$$r_{ks} = \sqrt{\sum_{n=1}^{N} (\frac{x_{ns}}{\overline{x}_n} - \frac{x_{nk}}{\overline{x}_n})^2 + \sum_{m=1}^{M} (\frac{y_{ms}}{\overline{y}_m} - \frac{y_{mk}}{\overline{y}_m})^2} \tag{3}$$

where r_{ks} is the radius of the sphere. $\overline{x}_n$ and $\overline{y}_m$ denotes the mean of inputs and outputs.

If we let the radius of the sphere increase until it contains the inefficient observation *k* and the efficient observation *s,* we can define the reference unit for the inefficient unit *k* as:

Definition: The efficient observation *s* is a reference to the inefficient observation *k* if

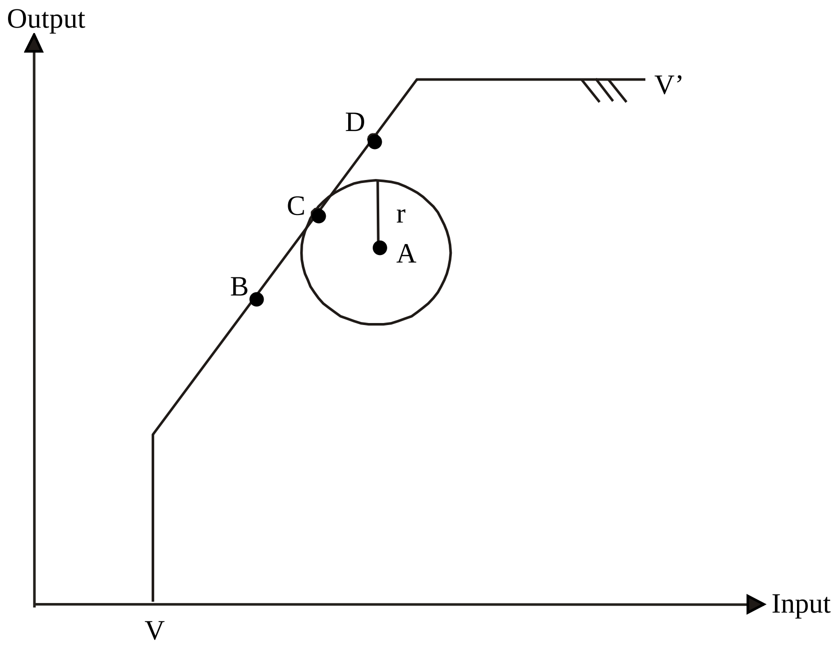
$$r_{ks'} = \text{minimum } r_{ks}, \ \forall s \in \mathbf{S}$$

min r_{ks} is thus the smallest distance between all efficient units and the evaluated

[13] It is possible that more than one reference unit exists. This will be the case if two, or more units have the value of the *sphere measure.*

inefficient unit *k*. The expression is interpreted as the minimum radius of the sphere, such that the sphere contains at least one efficient unit and the unit *k*. The solution to the minimising problem identifies the efficient unit that is located closest to the inefficient unit, measured by the Euclidean distance. The *sphere measure* is illustrated in Figure 2.

Figure 2. Illustration of the Sphere Measure



In this Figure, units *A*, *B*, *C* and *D* are the observed units, thus **K** = {*A*, *B*, *C*, *D*}. Among these units, *A* is inefficient, while *B*, *C*, *D* are efficient, thus **S** = {*B*, *C*, *D*}. When the radius, *r*, increases, unit *C* will be the first unit to appear within the sphere. The efficient unit *C* is then defined as a reference to unit *A*.[14] The result of the computation of the *sphere measure* for the data is presented in Table 6.

[14] Note that since the *sphere measure* searches for the most similar unit in all directions, it is possible that the selected reference unit use more input in one or more than one dimension. Depending on input prices this *could*, as in the intensity variable method, result in a situation of increased cost. To exclude this situation, information about input prices is necessary.

Table 6. Descriptive Statistics of the Sphere Measure

Inefficient unit no.	Mean r_{ks}	Std Dev. r_{ks}	Min. r_{ks}	Reference unit	Inefficient unit no.	Mean r_{ks}	Std Dev. r_{ks}	Min. r_{ks}	Reference unit
2	4.23	1.01	1.90	24	18	3.17	1.47	1.96	4
3	3.27	1.61	1.88	6	19	2.83	2.21	1.50	16
8	4.19	1.44	2.06	4	22	3.40	1.69	1.61	4
9	3.11	1.60	1.98	7	23	2.82	2.00	1.57	5
10	3.10	2.22	1.61	1	25	2.46	2.76	0.48	16
12	9.51	0.71	7.80	24	27	4.38	1.39	3.32	24
13	4.49	1.27	2.32	4	28	15.20	0.68	13.69	26
14	2.68	2.57	1.03	1	30	2.62	2.58	0.79	1
15	2.76	2.00	1.28	6					

Note: The Min. Radius represents the distance between the inefficient unit and the closest located efficient unit.

For the data used in this study, it was also possible to identify a unique reference unit with the *sphere measure*. Another appealing feature with the *sphere measure* is that a measure of proximity is also obtained. This makes it possible to evaluate the relevance of the identified reference unit. As can be seen from Table 6, the *sphere measure* varies from 0.48 to 13.69. This also indicates that some detected reference units are better suited than others.

V. Conclusions

The objective of this study has been to provide guidelines on what properties one can expect from a reference unit and also how these reference units could be detected. There is no doubt that reference units can play an important part when the results from an efficiency study are implemented in the investigated industry. Relevant reference units make it possible for an inefficient unit to study, on site, production that is more efficient than its own. This makes it possible for an inefficient unit to adopt a more efficient way to organise its production. The main question for this study has been how we can identify relevant reference units for a firm.

The literature suggested two methods, the *intensity variable method* and *the dominance method*. These methods were used on a data set on booking centre services in Sweden and some shortcomings were identified. Firstly, some pointed out reference units had little similarity with the inefficient unit. Secondly, when using the *dominance method*, no reference unit existed. These shortcomings motivate the search for a new method. To derive the new method, we started with a list of properties that are desired for a reference unit. A reference unit should always exist, the reference unit should be efficient, the reference unit should be an existing unit and finally, the reference unit should be similar to the inefficient unit. Given this list of properties; a new method labelled the *sphere measure* was developed. The idea with the *sphere measure* is to define a sphere around an inefficient unit and then expand the radius of the sphere until it contains the inefficient unit and at least one efficient unit. The unit that first appears in the sphere is then chosen as a reference unit. One advantage with the *sphere measure* is that it is constructed to fulfil all desired properties. In Table 7, the result concerning fulfilment of the four properties, with respect to methods are summarised.

By using the *sphere measure*, the efficient unit that has the largest similarity, measured by the Euclidean distance, is identified as a reference.

Table 7. Comparing Different Methods to Detect Reference Units

Property	Dominance	Intensity	Sphere
1 $\Re e_k \neq \emptyset$	No	Yes	Yes
2 If unit $j \in \Re e_k$ then $x_j \in Isoq\ (Ly)$	Yes	Yes	Yes
3 If unit $j \in \Re e_k$ then unit $j \in \mathbf{K}$	Yes	Yes	Yes
4 If unit $j \in \Re e_k$ then there cannot exist another unit i, $x_i \in Isoq\ (Ly)$, such that $\lVert ik \rVert < \lVert jk \rVert$	No	No	Yes

Appendix

Comparing the Input and the Output Vectors between Unit 7, Unit 11 and Unit 9

	Unit No. 9	Unit No. 11	Difference 9 vs.11	Unit No. 7	Difference 9 vs.7
Output					
Directly mediated services (Y1)	170,000	8,140	-161,860	150,000	-20,000
Co-ordinated and mediated services (Y2)	75,000	12,210	-62,790	150,000	75,000
Input					
Hours worked with booking - mediation (X1)	18,651	2,268	-16,383	5,017	-13,634
Hours worked with administration (X2)	1,049	0	-1,049	105	-944
Telephone lines to the booking centre (X3)	11	1	-10	5	-6
Floor space used by the booking services (X4)	55	9	-46	27	-28
Floor space used for administration (X5)	30	9	-21	10	-20
Value of purchased services in SEK (X6)	27,000	16,000	-11,000	70,000	43,000

VI. References

Althin, R., Färe, R. and J. Månsson (1994), *Beställningscentralers Kostnadseffektivitet: Effektivitet vid Produktion av Beställningscentraltjänster med Hänsyn tagen till Stödjande Organisation, (Efficiency in Production of Booking Centre Services, Including the Overhead Organisation)*, Kommunikationsforskningsberedningen, Stockholm, Sweden (In Swedish).

Charnes, A., Cooper, W.W and E. Rhodes (1978), “Measuring the Efficiency of Decision Making Units,” *European Journal of Operations Research* **2**: 429-444.

Färe, R., Grosskopf, S. and C.A.K. Lovell (1983), “The Structure of Technical Efficiency,” *Scandinavian Journal of Economics* **85**: 181-190.

Førsund, R.F. and L. Hjalmarsson (1974), “On the Measurement of Productive Efficiency,” *Scandinavian Journal of Economics* **76**: 141-154.

Førsund, R.F. and L. Hjalmarsson (1979), “Generalized Farrell Measure of Efficiency: An Application to Milk Processing in Swedish Dairy Plants,” *Economic Journal* **89**: 294-315.

Farrell, M.J. (1957), “The Measurement of Productive Efficiency,” *Journal of the Royal Statistical Society, Series A. General* **120** (Part 3): 253-281.

Kittelsen, S.A.C. and R.F. Førsund (1992), “Efficiency Analysis of Norwegian District Courts,” *Journal of Productivity Analysis* **3**: 277-306.

Månsson, J. (1996), “Technical Efficiency and Ownership: The Case of Booking Centres on the Swedish Taxi Market,” *Journal of Transport Economics & Policy* **30**: 83-93.

Tulkens, H. (1993), “On FDH Efficiency Analysis: Some Methodological Issues and Application to Retail Banking, Courts and Urban Transit,” *Journal of Productivity Analysis* **4**: 183-210.

von Neumann, J. (1938), “Über ein Ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes,” in K. Menger, ed., *Ergebuisse eines Mathematischen Seminars*, Vienna. Translated by Morgenstern (1945-46), *Review of Economics Studies* **13**: 1-9.

Journal of Applied Economics, Vol. VI, No. 1 (May 2003), 177-203

IMPORT PROTECTION, CAPITAL FLOWS, AND REAL EXCHANGE RATE DYNAMICS

Larry A. Sjaastad[*]

University of Chicago, and University of Western Australia

and

Meher Manzur[*]

Curtin University of Technology

Submitted October 2001; accepted August 2002

This paper focuses on the effect of import protection on the response of the real exchange rate to capital flows. The central hypothesis is that barriers to imports blunt the expenditure and production shifting effects of changes in relative prices, and hence the ability of the real exchange rate to equilibrate the economy in response to international capital flows. Employing a cross-section approach, the study focuses on three broadly similar countries but with very different levels of protection: Argentina, Australia, and Canada. The empirical results are consistent with the central hypothesis.

JEL classification codes: F13, F32, F41

Key words: import protection, real exchange rate

I. Introduction

This paper poses a connection between the level of protection of domestic

[*] Sjaastad (correponding author): lsjaasta@midway.uchicago.edu. The research reported in this paper was done at the University of Chicago, the Curtin University of Technology, and the Economic Research Centre of the University of Western Australia. The authors are indebted to participants at seminars at the Universities of Chicago, Miami, and Western Australia, the Curtin University of Technology, the Universidad Católica de Chile, the Universidad Nacional del Cuyo, and the Universidad del CEMA for helpful and constructive comments. The authors are particularly grateful to Victor Elias for providing much of the basic data for Argentina, and to Kenneth Clements for many helpful comments. Financial support from the Australian Research Council is gratefully acknowledged.

industry and the impact of international capital flows on the real exchange rate. The central proposition, one sketched out in an earlier paper by Sjaastad (1991), is that protection renders expenditure and production shifting between traded and home (i.e., nontraded) goods less responsive to relative prices, and hence increases the variance of the real exchange rate relative to that of capital flows; this occurs because protection reduces the volume of trade and, perhaps, the margins of substitution between traded and home goods as well. The result is that the real exchange rate reacts more strongly to capital flows in highly protected economies than in those with liberal commercial policies.

While it is obvious that import protection generates an import-competing sector unable to cope with foreign competition, it also has been found that an important manifestation of high protection is a retardation of industrialized exports (Miranda, 1986) and, consequently, an inordinate dependence on natural-resource-based export activities such as agriculture and mining. These industries are often slow in their ability to expand and contract, at least in the short run. In addition, as tariff structures are rarely uniform, imports become concentrated in low-tariff items which, in highly protected economies, tend to be capital goods, raw materials, and intermediate goods essential to the functioning of the protected industrial sector. This pattern of trade exacerbates the difficulty of adjusting to international capital flows; moreover, if the real exchange rate is rendered inflexible upwards by rigidity of both wages and the exchange rate, the necessary adjustments come about in quantities rather than prices, leading to the classic "stop-go" economy.

This paper sets out to test these ideas. In particular, we attempt to identify the effect of protection on the response of the real exchange rate to international capital flows, the central hypothesis being that, other things equal, protection leads to greater variability of the real exchange rate.

The remainder of the paper is organized as follows. Section II presents a selective review of the existing literature, and Section III develops a simple model that highlights the impact of protection on the behavior of the real exchange rate. The empirical methodology and results are presented in Section IV, in which estimates of the elasticity of the real exchange rate with respect to capital flows are found to be strongly affected by notional levels of protection. Policy implications are briefly discussed in the final section.

II. A Selective Survey of Existing Literature

The role of the real exchange rate in macroeconomic adjustment has become prominent in recent research on open economies such as that of Edwards (1988). It is typically argued that stable real exchange rates at appropriate levels send the correct signals to economic agents and facilitate smooth adjustment of the balance of payments, thereby ensuring macroeconomic stability and increased welfare; as Mussa (1982) has pointed out, however, the variance of purchasing power parity (PPP) real exchange rates, defined as ep^*/p, where e is the nominal exchange rate and p and p^* are the domestic and foreign price levels, respectively, has increased sharply since fixed parities among the major currencies were abandoned in 1973. It also is frequently argued that persistent deviations from PPP are often due to misguided government policies that influence the allocation of spending between traded and home goods and services.

A. The Salter Effect

Since Salter's seminal 1959 paper, it is widely accepted that real exchange rates respond to international capital flows, which have accelerated in recent years, particularly so in the developing countries over the past decade. The response of the real exchange rate to capital flows, however, appears to differ across regions. Sachs (1981) analyzed the linkage between real exchange rates and the current accounts in OECD countries and found that over the 1970s many of the deficit countries experienced real exchange rate appreciation, while surplus countries (which included Japan and the United States) showed real depreciation. Schadler (1994) finds that capital flows into Thailand, Spain, Mexico, Egypt, Colombia and Chile during the late 1980s and early 1990s lead to real appreciations, while the IMF (1991), Calvo et al. (1993), and Khan and Reinhart (1995) find that, on average, the Latin American countries experienced larger real appreciations than did the Asian countries.

One prominent explanation for these differences is that the two regions do not attract the same kind of capital; direct foreign investment was more important in Asia than in Latin America. Companies investing in a new plant are likely to import the necessary equipment to run it; as the capital inflows are used to pay for those imports, the real exchange rate is unaffected. Others argue that the Asian economies channel foreign capital into investment, whereas Latin Americans tend to spend it on consumption. A third argument is that Latin American central banks have been less successful in sterilizing capital inflows by open market operations; the efficacy of sterilization is, however, open to question as at best it is effective only in the short run. It is the purpose of this paper to provide a fourth explanation for these differences; namely, that the greater is the degree of openness of an economy, the weaker will be the response of the real exchange rate to capital flows.

B. Liberalization and the Real Exchange Rate: The Sequencing Issue

The behavior of the real exchange rate is highly relevant to the design of liberalization policies and their effect on the balance of payments; Khan and Zahler (1983) provide a systematic analysis of the short run effects of liberalization on both the current and capital accounts. Central to that issue is the proper sequencing of the liberalization of trade and capital movements; in this context, the “Southern Cone” syndrome is relevant. That syndrome refers to the Argentine, Chilean, and Uruguayan liberalization cum stabilization policies in the late 1970s and early 1980s. While in full pursuit of ambitious liberalization programs, all three countries adopted exchange-rate-based stabilization plans involving minute, pre-announced, and diminishing rates of devaluation of their currencies against the U.S. dollar – the infamous tablitas. This policy mix succeeded only partially in reducing inflation (partly because the dollar itself was rapidly depreciating until mid-1980), but did result in large capital inflows in response to sustained interest rate differentials.

By the early 1980s all three Southern Cone countries had experienced substantial real appreciations, and all were confronting severe balance of payments crises as well as deep recession. Fernandez (1985) argued that capital inflows played a fundamental role in the short run dynamics of the Argentine real exchange rate, an argument that has been echoed by Corbo (1985) in the Chilean context and by Hanson and De Mello (1985) for the Uruguayan case. It is noteworthy, however, that despite the similarity of their exchange rate policies, real appreciations were far larger in Argentina and Uruguay than in Chile, which may be in part due to commercial policy; as Bruno (1985) pointed out, an important contrast between Chile, on the one hand, and Argentina and Uruguay on the other, was the high and growing degree of openness of the Chilean economy.[1]

The Southern Cone experiences have been widely analyzed by Bruno (1985), Harberger (1982), McKinnon (1982), and Sjaastad (1983), among others and, while that literature offers significant lessons for economic policy, little systematic analysis has established a precise link between the degree of openness of the economy and the quantitative response of the real exchange rate to international capital movements – the central theme of this paper. Although many might agree with McKinnon (1982) on the danger of removing capital controls in the face of heavy protection, as well as with Bruno's (1985) argument that "one important lesson (from the Southern Cone) for the sequencing of markets would seem to be placing the current account far ahead of the capital account in terms of timing" (p. 868), a definitive analytical underpinning for these views is not evident. Some believe that, because asset prices can adjust instantaneously while prices of goods and services adjust gradually, the real exchange rate impacts more quickly and strongly on the capital account than the current account. Others, such as Frenkel (1983) in his two-horse carriage analogy, argue that the capital account adjusts more rapidly than does the current account. Unfortunately, this proposition is not a scientific one as it cannot be refuted empirically; since current account deficits, as measured, are identical with capital account surpluses (apart from errors and omissions), it is impossible to observe any difference in speeds of adjustment of the two accounts.

The contribution of this study to the sequencing issue lies in the evidence that protection magnifies the reaction of real exchange rates to capital flows with the implication that unless prices, wages, and/or the exchange rate are highly flexible, free movement of capital in the face of heavy protection may be a recipe for macroeconomic instability. This argument should not be interpreted as support of capital controls, but rather as a rationale for the view

[1] According to Fernandez (1985), from 1978 to 1981 the Argentine real exchange rate fell by 34 per cent, and De Mello et al. (1985) calculate the decline in Uruguay at nearly 46 per cent, whereas Galvez and Tybout (1985) estimate the Chilean real appreciation to have been only 20 per cent in the same period.

that the dismantling of those controls be held in abeyance until trade liberalization is largely complete.

III. Capital Flows and the Real Exchange Rate

This section presents a skeletal model that illuminates the link between real exchange rates and capital flows, and also examines the ways in which import protection can exacerbate the variability of the real exchange rate. In view of the evidence that PPP real exchange rates are subject to substantial measurement error (Sjaastad 1998a, 1998b), the real exchange rate in this study is defined as a price index for internationally-traded goods relative to an index for nontraded (or home) goods rather than the PPP version thereof.

A. A Model of the Home-Goods Sector

The relationship between capital flows and the real exchange rate is based on equilibrium in the market for home goods. The economy has three types of goods and services: importables, exportables, and home goods, whose price indices are p_M, p_X, and p_H, respectively. Under an exchange rate rule, p_H is endogenous, and with a money supply rule, p_M and p_X are endogenous; in both cases, the endogenous price(s) induces the requisite expenditure and production shifting to accommodate a capital flow. The supply of home goods, q_H^S, depends upon the three prices and gross domestic product (GDP), designated by g. The demand for home goods, q_H^D, is a function of the same three prices, GDP corrected for the terms of trade, designated by y, and capital flows, indicated by k; $k > 0$ implies a capital inflow. The actual capital-flow variable is $k_g = k/g$.

Letting upper-case letters be the natural logarithms of lower-case letters, a local log-linear version of the model can be written as follows:

$$\begin{cases} Q_H^S = constant + \varepsilon_{H,H} P_H + \varepsilon_{H,M} P_M + \varepsilon_{H,X} P_X + \varepsilon_{H,G} G \\ Q_H^D = constant + \eta_{H,H} P_H + \eta_{H,M} P_M + \eta_{H,X} P_X + \eta_{H,Y} Y + \eta_{H,k} ln(1 + k_g\ ) \\ Q_H^D = Q_H^S = Q_H \end{cases} \tag{1}$$

where $\varepsilon_{H,i} = \partial Q_H^S / \partial P_i$ and $\eta_{H,i} = \partial Q_H^D / \partial P_i$ for $i = H, M, X$; since the effect of the terms of trade on income is captured by Y, the $\eta_{H,i}$ elasticities involve only substitution effects. As both Q_H^S and Q_H^D are homogeneous of degree zero in the three prices, $\varepsilon_{H,H} + \varepsilon_{H,M} + \varepsilon_{H,X} = 0$, and $\eta_{H,H} + \eta_{H,M} + \eta_{H,X} = 0$. The parameter $\eta_{H,Y} = \partial Q_H^D / \partial Y = \left[P_H (\partial q_H^D / \partial y)\right] / \left[(P_H q_H^D)/y\right] = mps_{H,y} / aps_H$ is the ratio of marginal and average propensities to spend on home goods. As the parameter $\eta_{H,k}$ is the elasticity of q_H^D with respect to $1 + k_g$, the ratio of expenditure to GDP, we have $\eta_{H,k} = \partial Q_H^D / \partial \ln(1 + k_g) = (\partial Q_H^D / \partial k) / \left[\partial \ln(1 + k_g)/\partial k\right] = mps_{H,k} / aps_H^*$ in which $mps_{H,k} = p_H \partial q_H^D / \partial k$ and $aps_H^* = (p_H q_H^D)/(g + k)$; note that $\eta_{H,k}$ and $\eta_{H,Y}$ are not necessarily identical.

A local solution for P_H is the following:

$$P_H = constant + [\omega P_M + (1 - \omega) P_X] - \theta^* \ln(1 + k_g) - \gamma^* \ln(1 + TT) \qquad (2)$$

in which $\omega = (\varepsilon_{H,M} - \eta_{H,M})/(\eta_{H,H} - \varepsilon_{H,H})$ is the “shift” parameter in the theory of the incidence of protection (see Sjaastad, 1980), $\theta^* = \eta_{H,k} / (\eta_{H,H} - \varepsilon_{H,H}) < 0$, and $\gamma^* = \varepsilon_{H,g} / (\eta_{H,H} - \varepsilon_{H,H})$. Since changes in y and g have similar effects on Q_H^D and Q_H^S, those variables (and their parameters) were combined into a terms-of-trade variable TT, whose definition can be found in the Data Appendix.

From equation (2), $\omega = \partial P_H / \partial P_M = (\partial P_H / \partial P_T)(\partial P_T / \partial P_M)$, where P_T is a traded-goods price index. But the homogeneity postulate requires that $\partial P_H / \partial P_T = 1$, so it follows that P_T can satisfy that postulate if and only if $\partial P_T / \partial P_M = \omega$, a requirement that is met by defining P_T as $\omega P_M + (1 - \omega) P_X$. As the real exchange rate is defined (in natural logs) as $RER = P_T - P_H$, equation (2) is an implicit relationship between capital flows and the real exchange rate.[2] The explicit relationship can be written as:

$$RER = constant + \theta^* \ln(1 + k_g) - \gamma^* \ln(1 + TT) \qquad (3)$$

where θ^* and γ^* are the elasticities of the real exchange rate with respect to the expenditure-output ratio and the income effects associated with changes

[2] With $PT = \omega P_M + (1 - \omega) P_X$, it follows that $\partial RER / \partial P_M = \omega - \partial P_H / \partial P_M = 0$ and $\partial RER / \partial P_X = (1 - \omega) - \partial P_H / \partial P_X = 0$, so the real exchange rate as defined in the text is invariant with respect to changes in P_M and P_X brought about by protectionist measures that do not involve first-order income effects. That property is not shared by PPP real exchange rates.

in the terms of trade, respectively. The effect of import protection on the magnitude of the parameter θ^* obviously is the focal point of the analysis.

B. Some Consequences of Protection

Import protection affects the magnitude of θ^* via a scale effect and perhaps also through a substitution effect. The scale effect arises because a protection-induced decline in the volume of trade magnifies the proportionate response of imports and exports to capital flows. When imports and exports are twenty-five to thirty per cent of GDP, a capital inflow of five percent of GDP can be accommodated with a relatively small increase in imports and/or a small reduction in exports. But when import protection has reduced the volume of imports and exports to, say, seven per cent of GDP, the required adjustments are relatively much larger. The scale effect is analogous to one of the sources of the recent external debt service problem in Argentina. While many commentators have pointed out that the Argentine external debt was not unduly large relative to her GDP, the fact that intense import protection in that country has severely contracted the volume of Argentine international trade with the result that, during 2001, interest payments on her external debt were equal to approximately fifty per cent of her export revenue.

Concerning the substitution effect, it is evident from casual observation that countries pursuing liberal trade policies have substantial domestic production of a rather broad set of importables and quite highly diversified exports, the outputs of which can readily expand or contract in response to changes in the real exchange rate. But the picture is very different in countries engaged in intense import substitution. In the first place, those countries typically adopt bi-modal tariff structures; protection granted to targeted industries usually is prohibitive (so the goods produced by those industries are no longer imported) while nontargeted imports face rather low tariffs.[3] As the number of targeted goods increases, the composition of imports undergoes a radical change; imports become concentrated in capital goods, raw materials, and intermediate goods, products that lack domestic substitutes

[3] For example, in 1975 the average tariff in Uruguay (a highly protectionist country) was 117 per cent, but tariff revenue was only ten per cent of the value of imports.

and which are used in roughly fixed proportions with value added in the protected industrial sector. In the limit, prohibitive tariffs are so pervasive that no domestically-produced goods are imported and no imported goods are produced domestically; in that case, any substitution between imports and home goods becomes limited to the final demand for the output of the protected industrial sector, thereby greatly weakening the expenditure and production-shifting effects induced by changes in the real exchange rate.

A similar phenomenon occurs in the export sector. Import protection is shifted onto the export sector in the form of an implicit export tax, the shifting being effected via increased costs (particularly wages) relative to output prices in the export sector (Sjaastad, 1980, Clements and Sjaastad, 1984). As protection grows, the implicit tax also increases and those export-oriented activities employing internally-mobile resources are the most vulnerable and the first to succumb (Miranda, 1986). When protection becomes intense, the only exports to survive are those in which sector-specific inputs (typically natural resources) account for a large part of total cost; those inputs have no alternative but to absorb the implicit tax. Sector-specific inputs are typically found in agriculture and mining, where supply elasticities are known to be low, at least in the short run. In many small countries (e.g., Chile and Australia), domestic demand for mineral products is minuscule relative to production, so the degree of substitution in consumption between those products and home goods is very small; in the case of agriculture, that substitution effect is limited as the demand for food products is price inelastic. Thus trade barriers also diminish substitution possibilities between home goods and exportables.

The nature of the scale and substitution effects can be illustrated further in the context of our model; one way involves transforming the denominator of the coefficient θ^*, $\eta_{H,H} - \varepsilon_{H,H}$, into cross elasticities. Differentiating the identity $q_H^D p_H + mp_M = q_H^S p_H + xp_X + k$ with respect to p_H, where m and x are the quantities of imports and exports, respectively, and holding k, p_M, and p_X constant results in: $q_H^D + p_H \partial q_H / \partial p_H + p_M \partial m / \partial p_H = q_H^S + p_H \partial q_H^S / \partial p_H + p_X \partial x / \partial p_H$. Setting $q_H^D = q_H^S = q_H$, this expression can be written in elasticity form as:

$$\eta_{H,H} - \varepsilon_{H,H} = \varepsilon_{X,H}\alpha_X - \eta_{M,H}\alpha_M \tag{4}$$

in which $\varepsilon_{X,\,H} = \partial X / \partial P_H < 0$ and $\eta_{M,\,H} = \partial M / \partial P_H > 0$ are the cross elasticities of export supply and import demand with respect to p_H, $\alpha_X = (xp_X) / (q_H p_H)$, and $\alpha_M = (mp_M) / (q_H p_H)$ are the ratios of exports and imports to expenditure on nontraded goods. Combining equation (4) with the definition of θ^* results in an alternative expression for that parameter:

$$\theta^* = \eta_{H,\,K} / (\varepsilon_{X,\,H}\,\alpha_X - \eta_{M,\,H}\,\alpha_M) \tag{5}$$

The scale effect associated with import protection is quite evident as that protection diminishes both α_X and α_M, thereby increasing the magnitude of θ^*.[4] The substitution effect associated with import protection would be reflected in a smaller magnitude of the cross elasticities $\eta_{M,\,H}$ and $\varepsilon_{X,\,H}$. The strength of the substitution effect, however, is ambiguous. In the case of imports, for example, $\eta_{M,\,H} = (\partial m / \partial P_H) / m$, and import protection has a negative effect on both $\partial m / \partial P_H$ and m. Accordingly, the nature of the effect on $\varepsilon_{X,\,H}$ and $\eta_{M,\,H}$ can be established only on the basis of empirical evidence. It is important to note that even if import protection were to have no effect on either $\varepsilon_{X,\,H}$ or $\eta_{M,\,H}$, it still can have a profound effect on $\eta_{H,\,H}$ and $\varepsilon_{H,\,H}$.

A second way to illustrate the scale and substitution effects is to derive the direct and indirect effects of a capital flow on the volume of imports. Holding p_M, p_X, GDP, and the terms of trade constant we have:

$$\begin{aligned} d(mp_M)/dk &= p_M\left[\partial m/\partial k + (\partial m/\partial P_H)(\partial P_H/\partial k)\right] \qquad (6)\\ &= mps_{M,k} + (mp_M)\eta_{M,H}\left[\partial P_H/\partial \ln(1+k_g)\right]\left[\partial \ln(1+k_g)/\partial k\right]\\ &= mps_{M,k} + (mp_M)\eta_{M,H}\theta^*/(g+k)\\ &= mps_{M,k} - aps_M\eta_{M,H}\theta^* \end{aligned}$$

where $mps_{M,\,k}$ is the marginal propensity to spend on importables with respect to a capital inflow and $aps_M = (mp_M) / (g + k)$ is the import ratio. As was pointed out above, while import protection has an ambiguous effect on the

[4] Exports decline because import protection involves an implicit tax on exports; for evidence on that issue, see Sjaastad (1980), Clements and Sjaastad (1984).

elasticity $\eta_{M,H}$, it clearly reduces the import ratio, aps_M, and probably $mps_{M,k}$ as well, which reduces the right hand side of equation (6). While import protection may affect the magnitude of $d(mp_M) / dk$, the direction of that effect is unclear. Accordingly, there is a strong presumption that import protection must increase the magnitude of θ^* to offset the decline it induces in the magnitudes of both aps_M and $mps_{M,k}$.

IV. Empirical Methodology and Results

To test the central hypothesis of this paper one might specify θ^* as a function of a protection-level variable and estimate that relationship with time series data; that approach, however, is unpromising as efforts to quantity protection have met with meager success. The average (or median) tariff can be meaningless, as tariffs in highly protectionist countries tend to be either prohibitively high or quite low.[5] The ratio of tariff revenue to imports cannot distinguish between low and high levels of protection; moreover, neither measure can detect non-tariff barriers. In view of these difficulties, it was decided to determine if the magnitude of θ^* differs systematically across three small, broadly similar countries, Argentina, Australia, and Canada, all of which have abundant natural resource endowments but very different commercial policies. Canadian markets have been very open to international trade in recent decades while Australia reputedly has been one of the most protectionist of the OECD club. Argentina's aggressive protection of her industrial sector is legendary; indeed, the uniform tariff equivalent of the Argentine tariff structure in the in the decade of the 1970s has been estimated at 98 per cent.[6]

The summary data for the three countries in Table 1 indicate that the degree of "openness" (the ratio of exports plus imports to GDP) during 1978-92 is

[5] Due to bi-modal tariff schedules, tariff revenue is often a very small fraction of the average (or median) tariff rate. As was noted earlier, in 1975 when Uruguay was a highly protectionist country, her average tariff was 117 per cent, but tariff revenue was only about ten per cent of imports.

[6] The uniform tariff equivalent is the uniform tariff that would result in the same volume of trade as does the actual tariff structure. The estimate of the uniform tariff equivalent for Argentina is from Sjaastad (1981).

Table 1. Summary Statistics for Three Small Economies: Period Averages, 1978-92

Country	Population (millions)	Real GDP (billions, 1985 U.S. dollars)	Openess (%)
Argentina	30.3	168.1	14.90
Australia	15.6	216.2	34.07
Canada	25.0	394.2	52.25

Source: Penn World Tables and World Bank STARS database.

highest for Canada and lowest for Argentina. Canada out traded Argentina by three and half times and Australia did so by more than two times. As Canada's GDP was more than twice that of Argentina, this ranking conflicts with the idea that trade is more important for a small economy than a larger one. While factors other than protection affect a country's trading activity, there can no doubt that at least part of the large but perverse differences in the trade volumes of these three countries arises from vastly differing degrees of import protection.

A. An Indirect Test

The first test of the proposition that import protection increases the magnitude of θ^* was an indirect one based on the response of imports to capital flows described in the previous section. To test that proposition, a discrete version of equation (6) was specified as follows:

$$\Delta(mp_M \; / \; g)_t = constant + \beta \Delta k_{g,\,t} + u_t \tag{7}$$

in which β corresponds to $d(mp_M) \; / \; dk$.

As θ^* is posited to be a function of the degree of import protection, the quarterly data samples for the three countries had to be chosen to reflect periods during which their commercial policies were quite stable. In the Argentine case, the sample begins with 1978:1 and ends with 1992:4, after which there was an attempt at trade liberalization in that country. In the case of Canada, the sample starts with 1971:1 and ends with 1994:3, prior to the implementation of NAFTA. For Australia, the sample period is 1977:3 to 1994:3. When estimates were made simultaneously for the three countries, the common sample period is that of Argentina. For details, see the Data Appendix.

Equation (7) was estimated simultaneously using quarterly data for the three countries by the RATS nonlinear system routine using White's (1980) robust standard error estimator (NSYS-ROB). As θ^* is posited to be a function both the relative volume of trade and its composition, the period was limited to 1978:1 to 1992:4 to avoid significant changes in commercial policy in any of the countries involved. The overall level of protection in those countries was quite stable from the middle to late 1970s to the early 1990s, but commercial policy in both Argentina and Australia became somewhat more liberal in the course of the 1990s. Descriptions and sources of the data appear in the Data Appendix.

The estimates of β in equation (7), summarized in panel A, Table 2, range from 0.44 to 0.51 and all three are highly significant.[7] While the largest estimate is for Canada, the estimates are not significantly different from one another as none of the equality restrictions, reported in panel B, Table 2, are rejected. When those restrictions are imposed, the estimate of β, reported in panel C, Table 2, is 0.46 with a t statistic of 11.82. These results could obtain only if the magnitude of the Argentine θ^* far exceeds that of both Australia and Canada.

These results can be used to illustrate the magnitude of the scale effect. From the definition of β, we can write $\theta^* = (mps_{M,k} - \beta) / (aps_M \, \eta_{M,H})$. Assuming that $mps_{M,k} = aps_M$, $\beta = 0.5$, and $\eta_{M,H} = 1$, then $\theta^* = 1 - 1/(2\, aps_M)$. If $aps_M = 1/3$, then $\theta^* = -0.5$; however, if the import ratio has been reduced to 1/12 by import protection (as in the case of Argentina), the magnitude of θ^* increases dramatically to -5.0**.**

B. Individual Country Estimates of Real Exchange Rate Elasticities

The second test of the effect of import protection on real exchange rate behavior involved estimation of equation (3). For this test, a proxy for the

[7] In making the estimates of β, serial correlation in the residuals was reduced by allowing one lag on the dependent variable. The estimates reported in Table 2 are of the long run values of β.

Table 2. Simultaneous NSYS-ROB Estimates of Equation (7): Argentina, Australia and Canada, 1978:1-92:4

A. Unrestricted Estimates of ß

Parameter	Estimate	t-statistic	P-value
β_{ARG}	0.4396	6.5959	0.0000
β_{AUS}	0.4495	5.2078	0.0000
β_{CAN}	0.5134	5.0455	0.0000

B. Chi-Square Equality Tests on Unrestricted Estimates of ß

Restrictions	χ^2 Statistic	P-value
$\beta_{ARG} = \beta_{AUS}$	0.0062	0.9370
$\beta_{ARG} = \beta_{CAN}$	0.3726	0.5416
$\beta_{AUS} = \beta_{CAN}$	0.1787	0.6725
All three	0.3776	0.8279

C. Restricted Estimate of ß

Parameter	Estimate	t-statistic	P-value
β	0.4567	11.8155	0.0000

D. Summary Statistics (Restricted Estimates)*

Country	R^2	SEE	D-W	Ljung-Box test	
				$Q_{(6)}$	P-value
Argentina	0.7268	0.0077	2.1703	1.6952	0.9455
Australia	0.6811	0.0067	1.9439	6.6733	0.3521
Canada	0.6198	0.0069	1.8541	5.5223	0.4788

Note: * The coefficients of determination were adjusted for degrees of freedom.

real exchange rate was developed, one that one that avoids the difficulties in constructing a home-goods price index, P_H. In short, that price index was replaced with the overall price level, $P = aps_H P_H + (1 - aps_H)P_i$. The resulting proxy for the real exchange rate, $RERP = P_T - P = aps_H RER$, differs from the real thing only by the factor of proportionality aps_H. With this alteration, equation (3) becomes:

$$RERP_t = constant + \theta \ln (1 + k_{g,t}) + \gamma \ln(1 + TT_t) + \upsilon_t \tag{8}$$

in which $\theta = asp_H\, \theta^*$ *and* $\gamma = asp_H\, \gamma^*$.

B.1. Sims Causality Tests

While the maintained hypothesis is that international capital flows "cause" the real exchange rate, it can be argued that a change in the real exchange can by itself induce an international capital flow. A spontaneous shift in demand away from traded towards nontraded goods, for example would increase the relative price of nontraded goods and might generate a current account surplus and hence a capital outflow, at least in the short run. Therefore, prior to estimating equation (8), the Sims procedure was used to test for causality.

The real exchange rate proxy, *RERP*, and the capital flow variable, $1 + k_g$, were pre-filtered to eliminate serial correlation. Six leads and lags on the independent variables were permitted in all cases, and the causality test was based on the joint significance of the leads.

The results of the Sims tests appear in Table 3. From panel A it is evident that the hypothesis that capital flows "cause" real exchange rates is not rejected for any country. Panel B, however, indicates that the reverse causality is rejected in every country.

B.2. Preliminary Estimates of Equation (8)

Since the real exchange rate may respond to capital flows and the terms of trade with lags, equation (8) was parameterized as follows:

$$A(L)RERP_t = constant + \Theta(L)\ln(1 + k_{g,t}) + \Gamma(L)\ln(1 + TT_t) + \nu_t \qquad (9)$$

where $A(L) = \sum_{i=0}^{M} a_i \mathrm{L}^{\mathrm{i}}$ is a polynomial of degree M in positive powers of the lag operator L, and likewise for $\Theta(L)$, whose degree is N, and $\Gamma(L)$. The final effect on *RERP* of a permanent shock to k_g is defined as $\theta = \Theta(1)/A(1)$.

Table 3. **Sims Causality Tests: Argentina, Australia, and Canada**

A. Tests if Capital Flows Cause Real Exchange Rates

Country	$\chi^2_{(6)}$ Statistic	P-value
Argentina	27.9734	0.0001
Australia	17.9843	0.0063
Canada	26.9296	0.0001

B. Tests if Real Exchange Rates Cause Capital Flows

Country	$\chi^2_{(6)}$ Statistic	P-value
Argentina	3.8841	0.6924
Australia	10.7189	0.0975
Canada	9.7558	0.1353

Preliminary OLS estimates of equation (9), with lags added until the sums of the polynomial coefficients stabilized, indicated that the joint restriction $A(1) = \Theta(1) = 0$ could not be rejected for any of the three countries; as a result, $\Theta(1)/A(1)$, the estimator of θ, is indeterminate. To deal with that problem, $A(L)$ was replaced with the identity $A(L) = (1 - L)\tilde{A}(L) + L^M A(1)$, and similarly for $\Theta(L)$; the degrees of the new polynomials $\tilde{A}(L)$ and $\Theta(L)$ are M-1 and N-1, and the kth coefficient of $\tilde{A}(L)$, for example, is $\tilde{a}_k = \sum_{i=0}^{k} a_i$. With $A(1)$ and $\Theta(1)$ restricted to zero, equation (9) becomes:

$$\tilde{A}(L)\Delta RERP_t = constant + \tilde{\Theta}(L)\Delta\ln(1 + k_{g,t}) + \Gamma(L)\ln(1 + TT_t) + v_t \qquad (10)$$

and the estimator of θ now is $\tilde{\Theta}(1)/\tilde{A}(1)$.

In the preliminary tests, the restriction $A(L) = (1 - L)$ also could not be rejected for any of the three countries, which implies $\tilde{A}(L) = 1$ and $\theta = \tilde{\Theta}(1)$. But since $\tilde{\Theta}(L) = (1-L)\tilde{\tilde{\Theta}}(L) + L^{N-1}\tilde{\Theta}(1) = (1-L)\tilde{\tilde{\Theta}}(L) + L^{N-1}\theta$, where $\tilde{\tilde{\Theta}}(L)$ is of degree N - 2, the final version of equation (8) is the following:

$$\Delta RERP_t = con + \tilde{\tilde{\Theta}}(L)\Delta^2 \ln(1+k_{g,t}) + \theta\Delta\ln(1+k_{g,t-N+1}) + \tag{11}$$

$$+\Gamma(L)\ln(1+TT_t) + v_t$$

Estimates of θ based on equation (11), with lagged variables as instruments, were made for each country by OLS using Hansen's (1982) generalized method of moments (OLS-GMM).[8] As will be seen, the differences in the estimates of the *θ's are* very substantial and consistent with the results reported in Table 2.

Argentina

The joint restrictions $A(1) = \Theta(1) = 1$ are not rejected (see panel A, Table 4); with those restrictions imposed, the OLS-GMM estimate of θ is -6.19 (see panel B, Table 4). That estimate is significant at the 0.00 per cent level, and is striking in economic terms: during the sample period a capital inflow of five per cent of Argentine GDP would inflate her CPI relative to traded-goods prices by more than thirty per cent!

Australia

The estimates for Australia were made in the same way as for Argentina, and are summarized in Table 4. With the zero-sum restrictions imposed on $A(1)$ and $\Theta(1)$, the standard error of estimate is only 2.2 per cent, and the OLS-GMM direct estimate of θ, -2.10, is significant at the 0.00 per cent level and is about one-third the magnitude of the corresponding estimate for Argentina.

[8] In none of the three cases were the estimates of θ sensitive to variations of plus and minus 0.2 in the value of ω used to construct P_T.

Table 4. OLS-GMM Estimates of Real Exchange Rate Elasticities (Equation 11)

A. Chi-Square Tests on Joint Restrictions

Country	Restrictions	$\chi^2_{(2)}$ Statistic	P-value
Argentina	A(1) = Θ(1) = 0	1.3909	0.4988
Australia	A(1) = Θ(1) = 0	0.5887	0.7450
Canada	A(1) = Θ(1) = 0	0.7548	0.6857

B. Restricted Elasticity Estimates

Country	Parameter	Estimate	t-statistic	P-value
Argentina	Θ	-6.1914	-20.7895	0.0000
Australia	Θ	-2.0996	-7.7314	0.0000
Canada	Θ	-0.6605	-2.7427	0.0061

C. Summary Statistics*

Country	R^2	SEE	D-W	Ljung-Box test	
				$Q_{(8)}$	P-value
Argentina	0.8737	0.1205	1.8688	5.6328	0.6883
Australia	0.9670	0.0219	1.5905	5.4219	0.7117
Canada	0.9679	0.0280	2.0873	6.3998	0.6025

Note: * The coefficients of determination were calculated on the basis of the variance of *RERP* and adjusted for degrees of freedom.

Canada

In the Canadian case the estimate of θ was made in the same way as for Argentina and Australia and the results appear in Table 4. With the $A(1)$ and Θ(1) zero-sum restrictions imposed, the estimate of θ is very small (one third that of Australia and about one tenth that of Argentina) but is significant at less than the one per cent level. Due to Canada's liberal commercial policy, capital flows are accommodated with very modest adjustments to her real exchange rate.[9]

B.3. Simultaneous Cross-Country Estimates

To test the significance of the differences in the estimates, the *θ's* were estimated simultaneously for all three countries by NSYS-ROB; the results appear in Table 5. The estimates for Australia and Canada differ somewhat from those reported in Table 4, but in view of the standard errors; the two sets of estimates are not inconsistent. While the estimate of θ for Canada is positive, it does not differ significantly from the estimate reported in Table 4. Tests on cross-country equality restrictions on the θ parameter are summarized in panel B, Table 5; all restrictions can be rejected at well below the one per cent level, which lends further support to the central hypothesis of this study.

C. Further Tests on the Argentine Case

In April 1991 Argentina drastically reformed both her exchange rate and monetary régimes. The peso was fixed against the U.S. dollar and became convertible, thereby eliminating all capital controls. Nonetheless, peso interest rates converged only slowly to dollar rates, which resulted in a large capital

[9] Referring back to the discussion in Section III.B, the point estimates of θ indicate that the substitution effect may also influence the impact of import protection on the behavior of the real exchange rate. Given the elasticities in equation (5), the magnitude of θ^* varies inversely with the "openness" ratio. That inverse for Argentina is 3.51 times that of Canada whereas the estimate of θ_{ARG} is 9.37 times θ_{CAN}, and the inverse for Australia is 1.53 times that of Canada, while the estimate of θ_{AUS} is 3.18 times θ_{CAN}, which appears to leave considerable room for the influence of the substitution effect. But as $\theta_i / \theta_j = (aps_{H,i} / aps_{H,j}) (\theta^*_i / \theta^*_j)$, the ratios θ_i / θ_j and θ^*_i / θ^*_j may not be identical, so the differences between the ratios of the inverses of the openness ratios and θ_i / θ_j ratios may be due to the possibility that protection increases the average propensity to spend on home goods. But as the Argentine propensity can hardly be triple that of Canada, nor can the Australian propensity be double that of Canada, import protection must reduce the scope forsubstitution between home and traded goods.

Table 5. Simultaneous NSYS-ROB Real Exchange Rate Elasticity Estimates (Equation 11): Argentina, Australia and Canada, 1978:1-92:4

A. Simultaneous Elasticity Estimates

Parameter	Estimate	t-statistic	P-value
θ_{ARG}	-6.1183	-3.8694	0.0001
θ_{AUS}	-1.7389	-2.5620	0.0104
θ_{CAN}	0.3634	0.6211	0.5346

B. Chi-Square Equality Tests on Elasticities

Restrictions	χ^2 Statistic	P-value
$\theta_{ARG} = \theta_{AUS}$	7.9587	0.0048
$\theta_{ARG} = \theta_{CAN}$	11.6239	0.0007
$\theta_{AUS} = \theta_{CAN}$	7.1551	0.0075
All three	12.1893	0.0023

C. Summary Statistics[*]

Country	R^2	SEE	D-W	Ljung-Box test	
				$Q_{(8)}$	P-value
Argentina	0.9018	0.1056	1.8736	4.7711	0.7817
Australia	0.9582	0.0198	1.6281	4.8060	0.7781
Canada	0.9941	0.0119	1.4767	9.6346	0.2916

Note: [*] See note in Table 4.

inflow, much of which is thought to be repatriation of foreign assets — the “Miami” dollars — by Argentine residents. The inflation moderated sharply but did not cease; from 1991:1 to 1993:1, consumer prices rose by 66 per cent, while the wholesale price index, which is heavily weighted with traded goods, rose by only 18 per cent. The Argentine post-reform inflation, which often has been attributed to inertia, clearly was concentrated in the home goods and services sector. Due to these developments, the Argentine case merits further analysis.

The degree to which the Argentine inflation following the régime change was due to large capital inflows was examined by analyzing the residuals (corrected to have a zero mean) of the OLS-GMM estimate of equation (11). Those residuals were regressed on dummy variables defined for each quarter of the 1990:1-92:4 period; the dummy variables were set to unity for the quarter in question and zero for all others, and their coefficients (which are the exact residuals for the quarters in question) and standard errors were estimated by OLS with a separate run for each quarter. The results, which appear in Table 6, indicate that the model performs even better after the régime change than before; the average residual was 11.67 per cent in the five quarters preceding the régime change versus 3.93 per cent for the seven quarters

Table 6. Real Exchange Rate Equation Residuals: Argentina, 1990:1-92:4*

Final quarter	k_g (%)	Residual	Standard error	t-statistic	P-value
1990:1	-3.85	-0.1892	0.1082	-1.7491	0.0860
1990:2	-6.25	-0.2160	0.1072	-2.0144	0.0490
1990:3	-2.62	-0.0455	0.1110	-0.4099	0.6835
1990:4	0.33	-0.1058	0.1103	-0.9597	0.3415
1991:1	-2.81	-0.0268	0.1111	-0.2415	0.8101
1991:2	-0.96	0.0149	0.1112	0.1340	0.8939
1991:3	3.01	0.0655	0.1108	0.5912	0.5568
1991:4	4.71	0.1315	0.1097	1.1984	0.2360
1992:1	2.92	-0.0191	0.1112	-0.1722	0.8640
1992:2	5.18	0.0063	0.1112	0.0565	0.9551
1992:3	5.38	-0.0197	0.1112	-0.1771	0.8601
1992:4	5.06	0.0180	0.1112	0.1619	0.8720

Note: * Based on the estimate of equation 11 for Argentina, summarized in Table 4.

beginning with 1991:2. Moreover, after the change in régime, only one residual exceeded ten per cent and none were significantly different from zero. Indeed, in 1992, despite a capital inflow of nearly five per cent of GDP, the residuals were very small. Finally, while it might appear that negative forecast errors are associated with capital outflows, that association is very weak, as only one of the twelve residuals is significant at the five per cent level. These results support the position that the Argentine post-reform inflation resulted from capital inflows rather than sheer inertia.

IV. Summary and Conclusions

This paper has analyzed the impact of import protection on the reaction of the real exchange rate to international capital flows. The maintained hypothesis is that import protection reduces the quantitative response of demand and production to changes in the real exchange rate. The empirical results strongly support that hypothesis. The evidence from three small countries, Argentina, Australia, and Canada, indicates that during the period from the late 1970s to the early 1990s the response of the real exchange rate to capital flows was extremely large for Argentina (highly protectionist by any standard), quite substantial for Australia (highly protectionist by OECD standards) but negligible for Canada (a relatively free trading country). Indeed, the point estimates reported in Table 4 indicate that a capital inflow of five per cent of GDP would increase the Argentine price level relative to the price of traded goods by 31 per cent, versus ten cent in Australia and only three per cent in Canada. Moreover, the responses in all three countries differed significantly at less than the one per cent level.

When neither the exchange rate nor the nominal wage is flexible, capital flows can result in severe macroeconomic instability; the Argentine situation of 1995-96 is a case in point. Owing to the Mexican crisis of late 1994, the capital flow into Argentina reversed but, as the Argentine exchange rate was fixed and the labor market exhibited little downward flexibility in nominal wages, the real exchange rate mechanism could not come into play and the result was a singular increase in unemployment.These results also provide an insight into the issue of the sequencing of liberalization in developing countries that was discussed in Section I. Eliminating capital controls prior to liberalizing trade will sooner or later lead to capital flows, and since protection magnifies the response of the real exchange rate to capital flows, those flows will require large adjustments in the relative price of home goods and wages. Although it is hard to make a convincing case that capital movements are inherently bad, the results of this study indicate that when a country imposes heavy restrictions on current account transactions, it will do well to impose restrictions on capital account transactions as well, a proposition that conforms to the general theory of the second best. Although relaxing restrictions on international flows of both capital and goods is widely viewed as desirable, this study suggests that capital controls should not be dismantled until the commercial account has been substantially opened.

Data Appendix

All data were quarterly for periods ranging from the 1970s to the early 1990s. Augmented Dickey-Fuller unit-root tests (not reported but available upon request) on the relevant variables for all three countries (with a trend for variables when in level form) showed that, with four lags, unit roots were rejected for all variables at the three per cent level and, when the variables were first differenced, unit roots were rejected for all variables at the one per cent level for all lags.

In all cases k_g was defined as a fraction of GDP. As a GDP deflator was unavailable for Argentina, the proxy for the real exchange rate was defined on the consumer price index, p_C, in all cases. The P_T variable was defined as a weighted average of P_M and P_X using the ω parameter as defined earlier.

The exact form of the final term in equation (2), which was represented by $\gamma^* \ln(1 + TT_t)$, is $(\eta_{H, Y} Y_t - \varepsilon_{H, G} G_t) / (\varepsilon_{H, H} - \eta_{H, H})$. By definition, $y_t = g_t(1 + TT_t)$, where TT_t is a first approximation of the terms-of-trade income effect as a fraction of real GDP and defined as $TT_t \equiv (x_{t-1} \Delta p^*_{X,t} - m_{t-1} \Delta p^*_{M,t}) / g^*_t$ in which a * superscript indicates that the variable has been deflated by p_c. In the case of exports, $(x_{t-1} \Delta p^*_{X,t}) / g^*_t = \left[(p_{C,t} x_{t-1}) / g_t \right] \Delta p^*_{X,t} = (xp_X)_{t-1} \Delta P^*_{X,t} / g_t$ and similarly for imports, so $TT_t = \left[(xp_X)_{t-1} \Delta P^*_{X,t} - (mp_M)_{t-1} \Delta P^*_{M,t} \right] / g_t$. Combining $Y_t = G_t + \ln(1 + TT_t)$ with the numerator of the exact form of the final term in equation (2) yields $\eta_{H, Y} Y_t - \varepsilon_{H, G} G_t = (\eta_{H, Y} - \varepsilon_{H, G}) G_t + \eta_{H, Y} \ln(1 + TT_t)$.

As variations in y and g have similar effects on q_H^D and q_H^S, respectively, the elasticities $\eta_{H,Y}$ and $\varepsilon_{H,G}$ are both positive and similar in magnitude, the term $(\eta_{H,Y} - \varepsilon_{H,G})G_t$ was ignored and hence $\gamma^* = \eta_{H,Y} / (\varepsilon_{H,H} - \eta_{H,H})$.

Argentina

Most Argentine data are from the FIEL database. The export and import price variables are the wholesale price index for agricultural products, which are Argentina's main export, and the wholesale import price index, respectively. The value of ω, 0.48, for constructing P_T is from Sjaastad (1981). Because of problems with the Argentine balance of payments data, net factor payments abroad were excluded from the capital-flow measure in the Argentine case. Those payments were excluded because, during the period in question, Argentina had a large (gross) external debt, but her private-sector foreign assets were smaller but of a similar order of magnitude. While service of the largely official external debt does appear in the service account of the Argentine balance of payments, it is widely believed that the earnings on privately-held foreign assets do not because those earnings were largely unrepatriated, and no imputation was made to the balance of payments for those earnings. Since the factor service account of the Argentine balance of payments grossly overstates actual net service of external debt during the sample period, capital flows in the Argentine case were defined as the deficit in merchandise and non-factor service trade.

Australia and Canada

Australian and Canadian data are from TIME SERIES DATA EXPRESS (EconData Pty Ltd of Australia). Import and export prices indices are identified in the database as IMPIPI and EXPIPI, respectively. For both countries, the capital flow variable was defined as the deficit in the goods and services account of their balance of payments as a fraction of GDP. The values of ω, 0.60 for Australia and 0.76 for Canada, for constructing the traded-goods price indices were obtained from a study reported in Sjaastad (1998b).

References

Bruno, M. (1985), "The Reforms and Macroeconomic Adjustments: Introduction," *World Development* **13**: 867-869.

Calvo, G.A., Leiderman, L., and C. Reinhart (1993), "Capital Flows and Real Exchange Rate Appreciation in Latin America: The Role of Real Factors," *IMF Staff Papers* **40**: 108-51.

Clements, K. and L. A. Sjaastad (1984), *How Protection Taxes Exporters*, London, Macmilan (Thames Essay for the Trade Policy Research Center).

Corbo, V. (1985), "Reforms and Macroeconomic Adjustments in Chile During 1974-84," *World Development* **13**: 893-916.

De Mello, J., Pascale, R., and J. Tybout (1985) "Microeconomic Adjustments in Uruguay During 1973-81: The Interplay of Real and Financial Shocks," *World Development* **13**: 995-1015.

Devereux, J., and M. Connolly (1994), "Commercial Policy, the Terms of Trade and the Real Exchange Rate Revisited," *Journal of Development Economics* **50**: 81-99.

Edwards, S. (1988) *Exchange Rate Misalignment in Developing Countries*, Baltimore and London, Johns Hopkins Press.

Fernandez, R.B. (1985), "The Expectations Management Approach to Stabilization in Argentina During 1976-82," *World Development* **13**: 871-92.

Frenkel, J.A. (1983), "Remarks on the Southern Cone," *IMF Staff Papers* **30**: 164-73.

Galvez, J., and J. Tybout (1985), "Microeconomic Adjustments in Chile During 1977-81: The Importance of Being a Grupo," *World Development* **13**: 969-994.

Hansen, H., and K. Juselius (1995), *CATS in RATS: Cointegrating Analysis of Time Series*, Evanston, IL., ESTIMA.

Hansen, L.P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica* **50**: 1029-54.

Hanson, J., and J. De Mello (1985), "External Shocks, Financial Reforms and Stabilization Attempts in Uruguay during 1974-83," *World Development* **13**: 917-939.

Harberger, A.C. (1982), "The Chilean Economy in the 1970s: Crisis, Stabilization, Liberalization, Reform," in K. Brunner and A. Meltzer, eds., *Economic Policy in a World of Change*, *Carnegie-Rochester Conference Series on Public Policy* **17**: 115-52.

International Monetary Fund (1991), "Determinants and Systematic

Consequences of International Capital Flows," Occasional Paper **77**, IMF.

Khan, M.S., and R. Zahler (1983), "The Macroeconomic Effects of Changes in Barriers to Trade and Capital Flows: A Simulation Analysis," *IMF Staff Papers* **30**: 223-82.

Khan, M.S., and C.M. Reinhart (1995), "Capital Flows in the APEC Region," Occasional Paper **122**, IMF.

Mc Kinnon, R. (1982), "The Order of Economic Liberalization: Lessons from Chile and Argentina," in K. Brunner and A. Meltzer, eds., *Economic Policy in a World of Change*, *Carnegie-Rochester Conference Series on Public Policy* **17**: 159-86.

Miranda, K. (1986), "Manufactured Export Performances in Developing Countries: A Sectoral Trade Model Approach," Ph.D. Dissertation, University of Chicago.

Mussa, M. (1982), "Government Policy and the Adjustment Process," in J. Bhagwati, ed., *Import Competition and Response*: 73-122, Chicago, University of Chicago Press.

Rodriguez, C.A. (1994), "The External Effects of Public Sector Deficits," in W. Easterly, C.A. Rodriguez and K. Schmidt-Hebbel, eds., *Public Sector Deficits and Macroeconomic Performance*: 79-97, New York, Oxford University Press (for the World Bank).

Sachs, J. (1981), "The Current Account and Macroeconomic Adjustment in the 1970s," *Brookings Papers on Economic Activity* **1**: 201-68.

Salter, W.E.G. (1959), "Internal and External Balance: The Role of Price and Expenditure Effects," *Economic Record* **35**: 226-38.

Schadler, S. (1994), "Surges in Capital Flows: Boom or Curse?" *Finance and Development* (March): 20-23.

Sjaastad, L.A. (1980), "Commercial Policy, 'True Tariffs' and Relative Prices," in J. Black and B. Hindley, eds., *Current Issues in Commercial Policy and Diplomacy*: 26-51, London, Macmilan.

Sjaastad, L.A. (1981), "La Reforma Arancelaria en Argentina: Implicancias y Consecuencias," *Documento de Trabajo* **27**, Buenos Aires, CEMA.

Sjaastad, L.A. (1983) "Failure of Economic Liberalism in the Cone of Latin America," *The World Economy* **6**: 5-26.

Sjaastad, L.A. (1991), "Debts, Deficits, and Foreign Trade," *Economic Papers* **10**: 64-75.