AN ECONOMETRIC ANALYSIS OF U.S. GDP - DEMOCRAT VS REPUBLICAN: WHO GETS YOUR VOTE?

NICOLA DUNNE

Senior Sophister

The American presidential elections, undoubtedly the most widely debated set of elections worldwide, could perhaps be described as a popularity contest as opposed to a battle of policy objectives. In this econometric analysis Nicola Dunne investigates the differing effects of Democrat or Republican leadership on US GDP by employing US fiscal spending and US labour force participation as explanatory variables. The results of her paper beg the question: if American citizens chose their Presidents purely on the basis of economic effectiveness, would the outcomes be different?

Introduction

> *'Economic policy can result from governmental inaction as well as governmental action.'*[1]

A link between politics and economics has been present in nations across the world for quite some time. In America, presidents have repeatedly emphasised the importance of improving their country's economy. In fact, the many efforts made by different presidents, including job creation and fiscal spending, tended to shape the country's economic performance. Jimmy Carter for instance:

> 'worked hard to combat the continuing economic woes of inflation and unemployment. By the end of his administration, he could claim an increase of nearly eight million jobs and a decrease in the budget deficit' (www.whitehouse.gov).

In the course of this report, a detailed analysis of the impact that

[1] President John F. Kennedy, New York: 12/10/1960

Democrat and Republican presidents have on U.S. GDP will be conducted. It will be of great interest to gain insight into this relationship as the U.S. presidential elections draw closer. This is particularly the case due to the state of the American economy at the present time, where fears of a recession being imminent are relayed more and more frequently. The analysis will initially involve an outline of the econometric approach adopted. This entails a description of the estimation technique, data set and regression model. Afterwards, the results obtained will be used to support or reject the hypothesis that the political affiliation of American presidents tends to have a distinctive impact on U.S. GDP. Finally, a set of investigative tests will be carried out to evaluate the strength of the model itself, from which, an overall conclusion of the success of this regression will be made.

Econometric Approach

Firstly, it is necessary to specify that this study utilises a time series data set and also the ordinary least squares (OLS) method of estimation. Through the *Microfit* programme, a set of population parameters will be estimated, from which a "line of best fit" will be derived.

Population Regression Model

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3D + u$$

Where:

Y =	Dependent Variable
X_s =	Explanatory or Independent Variables
b_s =	Regression Coefficients
D =	Dummy Variable
u =	Error or Disturbance Term

The regression is developed using annual figures from the time interval, 1959 to 2003, which should contain sufficient information as ten Presidents have been in office during this period. Twenty-five of these years have seen a Republican president in charge, leaving nineteen years having a president of Democrat affiliation. This provides quite an even basis for examination.

Specification of Variables

Dependent Variable *Y*

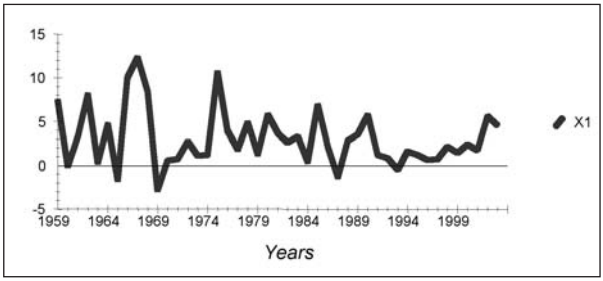
Y represents the dependent variable, which for the purpose of this report, is the growth rate in real U.S. GDP (2000 Prices) over the chosen period. The data for this variable was attained from the Economic Report of the President, 2006.

Figure 1. Plot of Growth Rate of GDP.

Independent Variables

X_1: The first independent variable is the growth rate of Annual Fiscal Spending (2000 Prices) by the U.S. government. The data for this variable was constructed through calculations using the U.S. Government's Budget figures, containing Historical Tables for the fiscal year 2007.

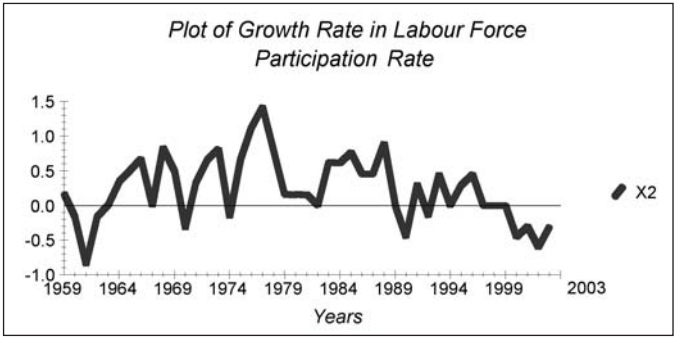
Figure 2. Plot of Growth Rate in Fiscal Outlays



X_2: The second explanatory variable chosen is the growth rate in the annual Labour Force Participation Rate in the U.S. economy. Again, the data for the

variable was composed through calculations using the Economic Report of the President, 2006.

Figure 3. Plot of Growth Rate in Labour Force Participation



Dummy Variable

D: To conduct this analysis, it is necessary to introduce a Dummy Variable 'as a device to classify data into mutually exclusive categories' (Gujarati, 2003: 298). In this case, Democrats are arbitrarily chosen as the benchmark category (0) and therefore Republicans become the alternative category (1). The information regarding the years when the various Presidents were in power and the political parties, of which each of these Presidents were members, is taken from web-based sources.[2]

Error Term

u; Due to the wide array of possible variables that may influence the level of GDP in a country, such as literacy levels or entrepreneurial spirit, it is necessary to incorporate an error term which represents all the omitted variables from the regression.

> 'No matter how many explanatory variables we include in our model, there will always be factors we cannot include, and these are collectively contained in u' (Wooldridge: 2006: 76).

[2] www.whitehouse.gov

Expectations

We would expect there to be a positive relationship between fiscal spending and GDP and also between the labour force participation rate and GDP as both are important stimulants for productivity, output and growth in an economy. One would also anticipate a negative relationship between the dummy variable and GDP due to the conservative aspect of the Republican Party. It can be argued that this conservative element would limit the party's ability to make strong economic improvements, relative to the Democrat Party. In addition, as this is a time series model, heteroskedasticity is not expected to be present in the analysis. However, due to the correlation patterns that tend to exist between some of the independent variables, we do expect to find some degree of multicollinearity.

Primary Results

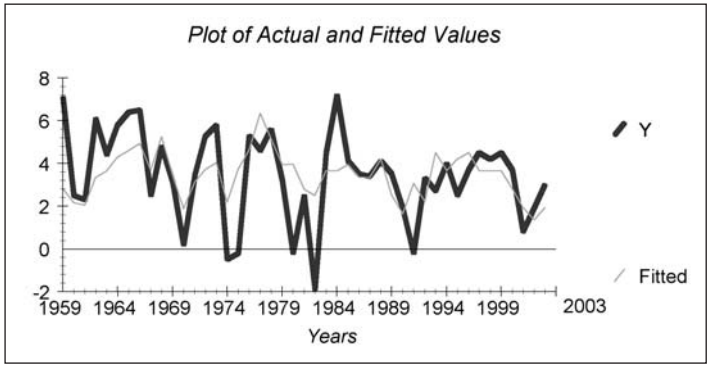
The results that the *Microfit* programme returned allowed us to construct the following line of best fit:

$$Y = 3.6507 + 0.0017043X_1 + 1.8851X_2 - 1.1671D$$

Table 1: Results

Explanatory Variables	Coefficients	T-Statistic	Standard Error	P-Value
C	*-3.6057*	*6.8457*	*0.53329*	*0*
X1	*0.001704*	*0.019625*	*0.086842*	*0.984*
X2	*1.8851*	*3.1417*	*0.60004*	*0.003*
D	*-1.1671*	*-2.0455*	*0.57055*	*0.047*

Figure 4. Plot of Actual and Fitted Values



Expectations versus Results

Table 2 demonstrates that the expected relationships between the dependent and independent variables in this model have been verified by the actual results. It is interesting to note that the dummy variable has a negative relationship with GDP, as predicted. This shows that, relatively speaking, Republicans are less successful at generating GDP growth, i.e. the average growth rate is 2.48% compared to 3.65% for the Democrat Party.

Table 2: Expectations and Results

Variables	**Expected Relationship**	**Results**
Y and X_1	Positive	Positive
Y and X_2	Positive	Positive
Y and D	Negative	Negative

Analysis of the Model

Now that the initial results have been reported, it is necessary to perform a number of different tests to examine this model more thoroughly.

R-Squared

R^2 is defined as 'the basic measure of goodness of fit in regression analysis'. In this regression, R^2 is found to be a low *0.25399* or 25.4%, meaning that just over 25% of the data is explained by the independent variables. This perhaps suggests that some of the variables omitted from the regression may have been more significant. Therefore, they should have been included.in order to obtain a greater level of fit in the analysis. At the same time, according to Achen, 'R^2 measures directly neither causal strength nor goodness of fit' (Achen, 1982: 64). Furthermore, it should not be strictly assumed that regressions are 'less satisfactory or less powerful if their R^2 is lower' (Achen, 1982: 59). Thus it is important to pay attention to the statistical significance of the explanatory variables and in particular to 'the underlying theoretical expectations about the model in terms of a priori signs of the coefficients entering the model' (Gujarati, 2003: 232).

T-Statistic

In this model, X_2 and *D* both have statistically significant partial effects on *Y* at the 5% level. Thus one can reject that the poplation parameters of these two variables have a true value equal to zero. However, X_1 has a very high p-value of *0.984*. As a result, it can be stated that this variable does not hold much significance.

F-Statistic

'In multiple regression models, the F-test is used to test the overall significance of the regression' (Brown, 1991: 108). The F-statistic, in this model, has been calculated as *4.6529*. Since the corresponding p-value is just *0.007*, I can reject the null hypothesis of all slope coefficients being simultaneously zero and thus assert that at least one of the explanatory variables has a significant effect on Y.

Autocorrelation

Autocorrelation can be detected through the use of numerous different tests such as the Durbin-Watson test. The DW statistic we obtain for this model is *1.6892*. Using a 5% significance level, given that there are 45 observations, with 3 explanatory variables, the upper and lower critical values are:

$$d_U = 1.615 \text{ and } d_L = 1.383 \text{ respectively.}$$

We fail to reject the null hypothesis of no autocorrelation when:

$$d_U < d < 4 - d_U$$

Therefore, since:

$$1.615 < 1.6892 < 2.3108,$$

the null hypothesis cannot be rejected and it can be concluded that no autocorrelation is present in this model. Therefore, the OLS estimators may still, at this point, be described as BLUE (Best, Linear and Unbiased Estimator). As a result, the Classical Normal Linear Regression Model (CNLRM) assumptions are not affected at this stage.

Functional Form

According to Gujarati, one of the assumptions under the CNLRM is that 'the regression model is correctly specified' (Gujarati, 2003: 73). This in part refers to the use of the correct functional form in the model, which can be analysed using a test known as Ramsey's Regression Specification Errors Test (RESET). The null hypothesis, stated as, correct model specification, should not be rejected if the F-statistic is significant at the 5% level. In this model, the F value is quite low at *0.69745* while the p-value of *0.404* is greater than *0.05*. This indicates that the model does not appear to be mis-specified.

Normality Test

The normality assumption tends to be employed as it results in the residual terms being normally distributed and in the case of small finite samples (such as this model) 'it enables us to use the t, F, and c2 statistical tests for regression models' (Gujarati, 2003: 110). The presence of normality can be examined using the Jarque-Bera test. The calculated statistic obtained in this test was *2.1766* with a p-value of *0.337*. According to Gujarati;

> 'if the computed p value of the JB statistic in an application is sufficiently low … one can reject the hypothesis that the residuals are normally distributed' (Gujarati, 2003: 148).

Due to the high p-value found, it is verified that the residuals are normally distributed and all of the benefits of this as highlighted above, can be exploited.

The histogram of residuals below seems to support this result. In addition, the plot of residuals within two standard error bands confirms that some residuals are positive while others are negative. This implies that the distribution of residuals is not skewed.

Figure 5. Histogram of Residuals and the Normal Density

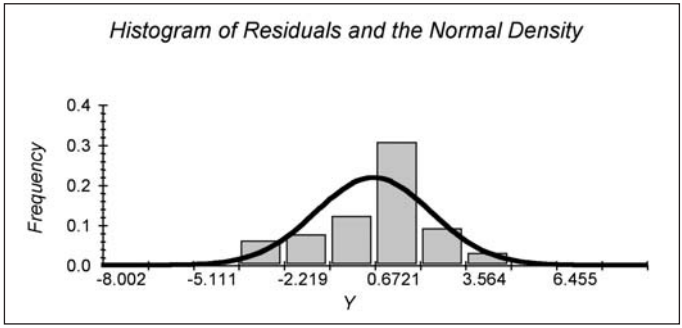
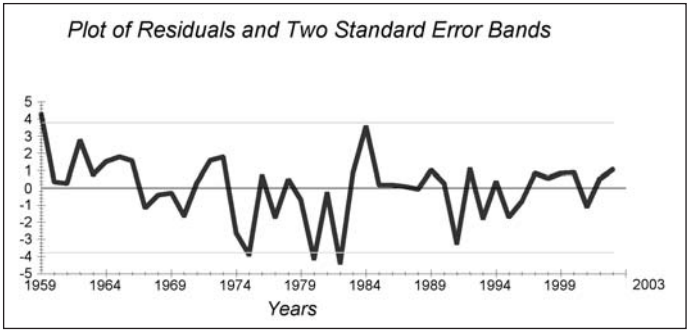


Figure 6. Plot of Residuals and Two Standard Error Bands



Heteroskedasticity

Microfit applies the Koenker-Basset test for heteroskedasticity. The null hypothesis of this test states that homoskedasticity is present. The computed F value in this model is found to be *0.26130* with a corresponding p-value of *0.612*. Due to the high p-value obtained, it is concluded that the null hypothesis should not be rejected. As expected, heteroskedasticity is not a problem that is faced by this model. Therefore, the residuals appearing in this regression are, 'homoscedastic; that is, they all have the same variance' (Gujarati, 2003: 387). As a result, this model's estimators maintain BLUE status.

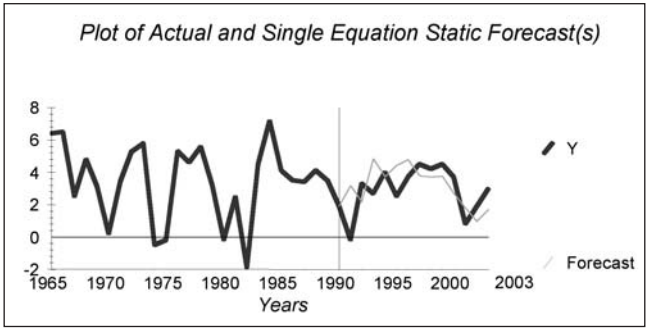
Multicollinearity

Multicollinearity, defined as, 'high (but not perfect) correlation between two or more independent variables' (Wooldridge, 2006: 102) is tested by regressing the different explanatory variables on each other, and checking the resulting R^2 that is found (Wooldridge, 2006: 102). In this model, a high level of multicollinearity exists between X_1 and X_2, where R^2 equals *0.91884*. This result was expected due to a common trend in the two variables i.e. both tend to increase over time. Although the OLS estimators will still be BLUE in the presence of multicollinearity, their variances will be effected. Large variances result in 'the t-ratio of one or more coefficients … [being] statistically insignificant' (Gujarati, 2003: 350). However, little correlation is found between either X_1 and D, or X_2 and D, as R^2 is found to be *0.024847* and *0.015410*, respectively.

Forecasting

The final point of reference, which is very important, is the forecasting ability of this model. According to Koutsoyiannis; 'Forecasting is one of the prime aims of econometric research' (Koutsoyiannis, 1977: 28). By eliminating 13 years from the original regression, it is possible to use the data collected to forecast the GDP growth rates for the eliminated years. As the following graph demonstrates, while the forecast is not perfect, it does provide a satisfactory estimation of the trends to be expected in GDP. This supports the significance of the variables used in this regression.

Figure 7. Plot of Actual and Single Equation Static Forecast(s)



Conclusion

This report provides a useful insight into the impact that Democrat and Republican presidents have on U.S. GDP. Given the onslaught of negative shocks to the American economy in recent times, namely the sub-prime mortgage crisis, the consequent credit crunch and the significant decline in the housing market, a considered view of the economic effectiveness of Democrat and Republican presidents is all the more relevant as we now approach elections in November of this year. The results of this investigation reveal that Democrat presidents have created a 3.65% average GDP growth rate, compared to a 2.48% average growth rate generated by Republican presidents. An important and statistically significant explanation for this difference may be the presidents' ability to create jobs in the economy, a variable that has a positive impact on GDP growth. However, the growth rate of fiscal outlays was proven to be statistically insignificant.

While some weaknesses such as a low R^2 and high multicollinearity are evident in this model, the data is strong in terms of autocorrelation, functional form and normality. Therefore, the result that Democrat presidents have created higher GDP growth levels, relative to their Republican counterparts, should not be disregarded. Consequently, a reasonable prediction, should a Democrat be elected president next term, is that an increase in GDP growth will occur. As a result, perhaps for the sake of the U.S. economy alone, the most useful outcome in November would be the arrival of a Democrat politician to the White House. It will be exciting to witness events unfold and to see if this will in fact come to pass.

Bibliography

Achen, C. H, 1982. *Interpreting and Using Regression*. United States of America: Sage Publications, Inc.

Brown, W, 1991. *Introducing Econometrics*. St. Paul, United States: West Publishing Company

Gujarati, D, 2003. *Basic Econometrics*. New York: McGraw Hill Companies

Koutsoyiannis, A, 1977. *Theory of Econometrics-An Introductory Exposition of Econometric Methods*. London: Macmillan

Wooldridge, J, 2006. *Introductory Econometrics*. USA: Thompson

Data Sources

http://www.whitehouse.gov/history/presidents/

http://a257.g.akamaitech.net/7/257/2422/15feb20061000/www.gpoaccess.gov/eop/2006/B39.xls

http://www.whitehouse.gov/omb/budget/fy2007/sheets/hist01z3.xls

http://www.whitehouse.gov/history/presidents/jc39.html

http://a257.g.akamaitech.net/7/257/2422/15feb20061000/www.gpoaccess.gov/eop/2006/B4.xls

AN ECONOMETRIC INVESTIGATION INTO THE DETERMINANTS OF INFANT MORTALITY RATES

NIALL SHERRY

Junior Sophister

Given the unprecedented worldwide economic progress of the past century, the fact that premature death amongst infants is still widespread constitutes one of the greatest humanitarian travesties of our time. In this paper Niall Sherry interestingly points out, that the wealth of a country is not necessarily a safeguard against this issue. Examining the relationship between various developmental indicators and their influence on infant mortality rates across a broad range of countries, his results suggest that while income growth is a determining factor, short-term medical relief may in fact be more effective than long-term policy measures.

Introduction & Motivation

'I believe that children are our future…'
Whitney Houston, 'The Greatest Love of All'

It is quite evident today that high infant mortality rates are a major problem in the developing world. However, there is also scope for investigation into the determinants of disparate infant mortality rates across the developed world, where economic advancement does not automatically bring about a corresponding drop in mortality rates. Even within some of the wealthiest nations in the world, there exists a relatively large disparity in infant mortality levels. These differences can be clearly seen when one compares Qatar (21 per 1000) with Japan (4 per 1000); or Saudi Arabia (26 per 1000) with Singapore (3 per 1000). Even within highly developed countries like the USA, infant mortality rates can at times remain high relative to rates in poorer nations. This project aims to measure infant mortality rates across the world by considering a number of variables linked to income, healthcare, equality and female education in order to determine whether causal relationships can be established. A thorough understanding of the causes of infant mortality can be of use to governments in both the developed and developing

world as they strive to reduce mortality rates. Such knowledge can potentially help ensure that governments of developing nations achieve a more equitable distribution of resources. This paper will also analyse how relevant these causes are at particular stages of economic development by running the overall model on specific subsets of the data.

Theory and Variables

The dependent variable (LINFMOR) for this project was taken to be the infant mortality rate, specifically the rate of neonatal death in children under one year of age per 1,000 live births, as measured by the World Bank.

The first independent variable (LINCOMEQ) to be included in the model was average income per person in each country as measured by GNI per capita. This approach follows the Atlas method. The variable acts as a general indicator of the economic status of the country. Here, we expect GNI per capita to have a negative impact on the mortality rate. Interestingly, past research may suggest that this relationship is not intuitive. From his study of infant mortality rates in England over the 16th, 17th and 18th centuries, Schellekens points out that: 'Since a rise in living standards may have improved parents' ability to acquire breast-milk substitutes, post-neonatal mortality could have risen with living standards', as breast milk provides a vital boost to the immune system of newborn infants (Schellekens, 2001: 4). However, the results of his research show a negative correlation between infant mortality and economic development. Schellekens suggests a number of reasons for this. For instance, he asserts that the eighteenth century decline in the neonatal mortality rate was: 'due to an improvement in the health and nutrition of mothers' (Schellekens, 2001: 9).

The second independent variable (LHEALTHQ) was the level of governmental expenditure per capita on healthcare. Again, a negative relationship was expected here, since more government spending in the area of healthcare should logically lead to better post-natal treatment for infants and the better provision of medicine. However, Andes stresses the importance of providing not just more healthcare, but good healthcare; consequently a third independent variable (DIAR) was included to measure the effectiveness of treatment within the given countries (Andes, 1989). Specifically, this variable refers to the percentage of children who receive fluids and treatment when suffering from diarrhoea (which was selected over immunisation rates for DPT and measles based on their poorer performance in the model).We would anticipate that higher percentages of treatment would lead to lower mortality rates. In practice, however, the latter of these two ended up being the sole

measure of healthcare effectiveness in the sample countries. The reasons for this are detailed fully below.

The fourth independent variable (FEMLIT) was female literacy rates. One of the most important factors in preventing infant mortality is well educated mothers. Waldmann posits that:

> 'greater female literacy could reduce infant mortality because literate women are more likely to know how to protect their infants' lives…Alternatively, high female literacy rates may indicate an egalitarian government devoted to primary education for all, which also places a high priority on public health' (Waldmann, 1992: 1294).

Glewwe, who specifically investigated the relationship between the two in Morocco, concurs with the reasoning that:

> 'Health knowledge appears to be the most important skill that mothers (indirectly) obtain from their schooling that prepares them to provide for their children's health' (Glewwe, 1999: 154).

He also provides data to demonstrate the relationship. The literacy rate of women aged 15-24, which is the primary age group for new mothers in most countries around the world, was used here as a measure of their human capital. This indicator was selected over the female primary education rate due to its better performance within the model. The literacy rate of women was expected to have a negative partial effect on the infant mortality rate.

Finally, inequality within countries needed to be examined for two reasons. First, to again quote Andes: 'a comprehensive account of the social processes important in reducing infant mortality includes an understanding of the socioeconomic context of the community' (Andes, 1989: 395). Inequality is one way of measuring the general social context of a country. Secondly, inequality has proven to be a strong influence on infant mortality in the past. Although Waldmann cannot conclusively offer any theory as to why this is, he fails to reject:

> 'the possibility that the results are caused by a positive correlation of the rich share and the relative price of e.g. health care…[or] the possibility that a larger fraction of babies are born to poor families in countries with a high rich share' (Waldmann, 1992: 1299).

Irrespective of the reasoning, inequality within countries has been demonstrated in the past to be a good determinant of the infant mortality rate. The findings below add some more weight to the theory, even if discussing or demonstrating any further reasoning or causality is beyond the scope of this project. To measure the effect of inequality, the Gini coefficient of income inequality (GINI) is used. This is the only independent variable to be included for which a positive parameter sign was expected, since as inequality rises, one would expect the infant mortality rate to rise as well.

Data

By and large, the data used in this project comes from the World Bank's database for the year 2000. This year was chosen as it contained the most complete and sufficient datasets for the largest number of countries. This allowed for complete data to be collected on 100 countries in total. However, several points should be noted.

For certain observations, in lieu of any data for a country in a given year, data from several years to either side of 2000 was used. It was anticipated that variables such as the female literacy rate wouldn't change too much within a short period of time. Thus, the results drawn from this project are still, by and large, statistically significant.

For two of the independent variables, diarrhoea treatment and literacy rates, data on first world countries wasn't available for some observations (due to the fact that a lack of diarrhoea treatment is not a problem in, say, the United States). In order to attain a feasible dataset, these were assumed for a number of first world countries to be virtually total, and set at 99%. While this somewhat diminishes the accuracy of the model, any discrepancy between the real data and the substitute data is likely to be minuscule and therefore insignificant.

For one independent variable, the Gini coefficient, the data was taken not from the World Bank but from the United Nations development programme as more data was available from the latter. Also of note is the large discrepancy in collection years for the data used in calculating these coefficients, ranging from the early 1990s up to 2007. Once again, while the limiting effect this divergence has on our ability to draw inferences from the data below is acknowledged, it is again contended that large-scale income inequality within a country should remain relatively constant over such a short time period.

Empirical Results

The population regression model that was utilised in this investigation took the following functional form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$$

where:

Y = Infant mortality rates per 1,000 live births
X_1 = GNI per capita, measured by the Atlas method
X_2 = Percentage of diarrhoea treatment
X_3 = Female literacy rates between 15-24
X_4 = The Gini coefficient of income inequality
u = The error term of statistical residuals

Note: the exogenous variable, government health expenditure, has been omitted from the model.

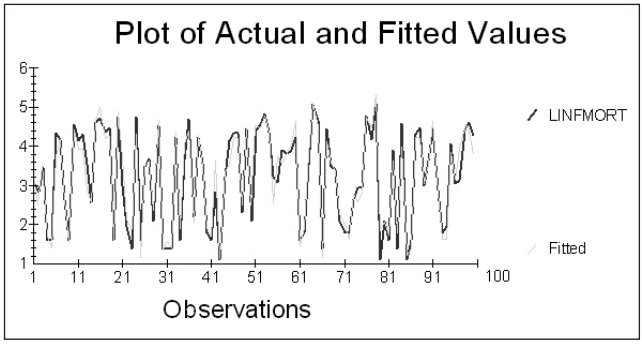
In this study, it was deemed appropriate to convert the endogenous variable as well as the X_1 and X_4 exogenous variables to log form in order to reduce heteroskedasticity and obtain a better functional form. Furthermore, the X_1 variable was also transformed to quadratic form. This allowed for a better reflection of the fact that the effect of increased income on infant mortality diminishes as income rises (as observed by running the model through *Microfit* in various forms). Once the data was passed through *Microfit* the subsequent set of results were attained.

Table 1: Regression Results

Regressor	Coefficient	Standard Error	T-Ratio	P-Value
INTERCEPT	3.8229	0.58226	6.5656	0.000
LINCOMEQ (X1)	0.20175	0.014517	-13.8975	0.000
DIAR (X2)	-0.010081	0.0015443	-6.527	0.000
FEMLIT (X3)	-0.006973	0.0018005	-3.8728	0.000
LINEQ (X4)	1.0040	0.14505	6.9213	0.000
R-Squared	0.94114			
R Bar-Squared	0.93866			
F-Statistic F(4,95)			379.7312	0.000
Functional Form CHSQ (1)			0.41333	0.520
Normality CHSQ(2)			2.9904	0.224
Heteroskedasticity CHSQ(3)			0.066697	0.796

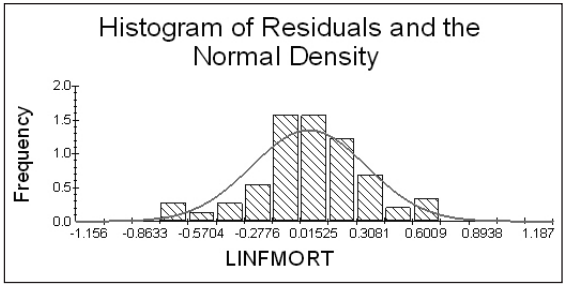
The R-squared result of 94.1% describes how much of the variation in the Y variable is explained by the model. Generally, it is prudent to also consider the R Bar-Squared result when a model consists of many exogenous variables. This result essentially imposes a penalty for adding extra independent variables to the regression. However, only four regressors are included in the model above and thus we would expect the two figures to be similar. A quick glance at the regression output table confirms this. An analysis of the t-ratios indicates that all the population parameters are statistically significant at the 1% level. This is also true of the intercept term. Alternatively, the same results can be inferred by looking at the p-values. Small p-values usually reflect evidence against the null hypothesis (H_0: $\beta_j = 0$). Furthermore, all of the signs for the coefficients fall in line with expectations. Given the large calculated F-statistic value of 379.7, we reject the null hypothesis that this model has no explanatory power. In fact the p-value suggests that there is no significance level at which it should not be rejected. As well as this, a plot of the fitted vs. actual values of Y implies that the sample regression function fits the data well.

Figure 1. Plot of Actual and Fitted Values



The heteroskedasticity result is also promising. Using the Koenker-Bassett test for heteroskedasticity, we fail to reject the null of homoskedasticity even.at the 79.6% significance level. Since the data used in this model is cross sectional, autocorrelation between residuals is not expected. Therefore, the results of the Durbin-Watson and Breusch-Godfrey tests are ignored. Ramsey's RESET test for correct functional form produces a small chi-square value and quite a high p-value. Thus we fail to reject the null hypothesis of correct functional form at any of the regular significance levels (1%, 5%, 10%). The Jacque-Berra test of normality of residuals reports a chi-square value of almost 3 and a p-value of 22.4%. Such figures result in a failure to reject the null of normally distributed residuals at the 10% significance level. In addition, the histogram of residuals seems to support the conclusion of the JB test. This solidifies the belief that the disturbances in the model follow a normal distribution.

Figure 2. Histogram of Residuals



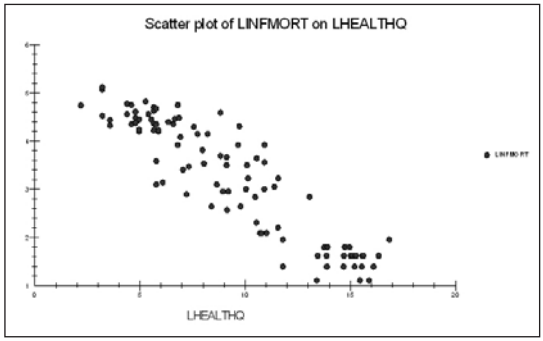
It is interesting to note at this point why one of the original variables, government health expenditure, has been excluded. The following are the regression results when it is included:

Table 2. Regression Results Including the Exogenous Variable LHEALTHQ

Regressor	Coefficient	Standard Error	T-Ratio	P-Value
INTERCEPT	3.9710	0.80383	4.9401	0.000
LINCOMEQ	-0.21942	0.06735	-3.2580	0.002
LHEALTH	0.01616	0.060120	0.26879	0.789
DIAR	-0.010167	0.0015847	-6.4156	0.000
FEMLIT	-0.0070433	0.0018282	-3.8527	0.000
LINEQ	0.99812	6.7724	6.7724	0.000
R-Squared	0.94118			
R Bar-Squared	0.93805			

As can be seen, the result here for health expenditure is insignificant at all of the usual levels. It adds almost nothing to the explanatory power of the model. However, this is not the truly worrying part. This model suggests that health expenditure has a positive partial effect on infant mortality. That is, the more money the government spends on healthcare, the more likely newborn babies are to die. Aside from being intuitively ridiculous, this is clearly proven wrong by examining the relationship purely between health expenditure and mortality (see Figure 3).

Figure 3. Scatter Plot of LINFORM on LHEALTHQ



We can clearly see a negative relationship between the two variables. Indeed, when the model is constructed with healthcare but without income included among the independent variables, the result is not dissimilar from the original model. In this case, healthcare becomes the most important variable with a strong negative relationship between it and the dependent variable. The reason for this effect is the incredibly strong correlation of .9683 between average income and government healthcare expenditure, close to perfect positive correlation.

This suggests an interesting effect whereby, the world over, government spending on welfare is completely dependent on the general welfare of the country irrespective of the type of government or party currently in power. However, a further investigation of this issue would go beyond the scope of this project.

Before we move on to analyse the results themselves, it is also worth including here the results of two sub-tests that were performed on the richest 25 and poorest 25 countries included in the original data set.

Table 3. Rich Countries Regression Results

Regressor	Coefficient	Standard Error	T-Ratio	P-Value
INTERCEPT	3.9296	21.4031	0.01836	0.856
LINCOMEQ	-0.04287	0.046487	-0.92221	0.367
DIAR	0.013004	0.21451	0.60620	0.551
FEMLIT	-0.2012	0.057522	-3.4977	0.002
LINEQ	1.637	0.24662	6.6379	0.000
R-Squared	0.83089			
R Bar-Squared	0.79707			

Table 4. Poor Countries Regression Results

Regressor	Coefficient	Standard Error	T-Ratio	P-Value
INTERCEPT	4.2911	1.6990	2.5256	0.020
LINCOMEQ	-0.17572	0.092475	-1.9002	0.072
DIAR	-0.007983	0.005481	-1.4566	0.161
FEMLIT	-0.009295	0.002563	-3.6266	0.002
LINEQ	0.80172	0.32445	2.710	0.023
R-Squared	0.70070			
R Bar-Squared	0.64084			

Analysis & Conclusions

Using the original data set, the obvious conclusion is that income is the most important determinant of infant mortality. This makes perfect sense. More money in a country and in the average family means more money available to keep children alive. Taking the high correlation between income and healthcare expenditure discussed earlier into account allows us to also conclude that the most powerful way to prevent infant mortality is to promote economic development. Although this is a simplistic and perhaps obvious result, it is an interesting one from an economic point of view. Also worth noting is the strong emphasis placed on the distribution of wealth as equally as possible, and the necessity of providing basic healthcare services such as diarrhoea treatments to children. This last one is perhaps the most interesting in terms of immediate action as it suggests that swift provision of basic medical care is more important than long-term education and improvement programmes.

However, once we start to analyse the separate sub-data sets for rich and poor countries, the results change slightly. For the 'rich countries' sub-set, income has become an almost irrelevant factor. As mentioned in the introduction, this can be seen empirically in the contrasting mortality rates of the various rich countries. Instead, by far the most important factor in developed countries is the distribution of wealth. As noted previously, it may be impossible to determine the exact causal link in this relationship but there does appear to be a very clear and strong correlation between the two. Interestingly, female literacy

is still a relatively important measure, suggesting that even countries in the developed world have some way to go in terms of educating their populations equally. Less surprising, however, is the fact that the population parameter for the exogenous variable, diarrhoea treatments, has now become statistically insignificant at the 10% level. These are relatively standard across all first world countries, and the question of basic healthcare provisions is a much less pressing one than in poorer countries.

Indeed, looking at the 'poor countries' sub-set, we can see that the population parameter for diarrhoea treatments is surprisingly insignificant at the 10% level, while education rates for females still have a statistically significant effect on the infant mortality rate. Income and inequality play a vital role in these areas as the internal distribution of wealth is often imbalanced. The differences between the various countries are often more pronounced than in richer countries due to the greater marginal value of a dollar in poorer nations. This is why the income variable entered the model in quadratic form. Also worth noting is the drop in the R-squared and R bar-squared figures to 70% and 64% respectively. This suggests, unsurprisingly, that there are more variables involved in dealing with less developed countries such as the availability of clean water.

Thus, our conclusions are somewhat perplexing. The original model seems to lack at least some explanatory power due to the widely varying socioeconomic nature of the countries being observed. The sub-sets, however, are too small to be of any real statistical value. Rather they offer mere suggestions of what could be uncovered by using additional observations in a more focused way. Therefore, the aim of this project was not to offer definitive results, but instead to propose suggestions. The first of these is that, at the most basic level, economic growth is the best way to prevent infant mortality. The second is that the distribution of economic wealth within a country defines the general socioeconomic outlook of that nation and thus is immensely important. Finally, immediate medical provision appears to be a significant factor in preventing infant mortality and should perhaps be the focus of those nations aiming to deal with the problem in the short-term. The problem of infant mortality will not be resolved overnight, but well targeted short-term solutions can help to prevent immediate loss and harm.

Bibliography

Andes, N. 1989. 'Socioeconomic, Medical Care, and Public Health Contexts Affecting Infant Mortality: A Study of Community-Level Differentials in Peru'. *Journal of Health and Social Behaviour* 30:4:386-397.

Glewwe, P. 1999. 'Why Does Mother's Schooling Raise Child Health in Developing Countries? Evidence from Morocco'. *The Journal of Human Resources* 34:1:124-159.

Schellekens, J. 2001. 'Economic Change and Infant Mortality in England, 1580-1837'. *Journal of Interdisciplinary History* 32:1:1-13.

Waldmann, R. 1992. 'Income Distribution and Infant Mortality'. *The Quarterly Journal of Economics* 107:4:1283-1302.

Data drawn from the World Bank website: (http://devdata.worldbank.org/dataonline/) and the UN Human Development Report: (http://hdrstats.undp.org/indicators/147.html)

AN INVESTIGATION INTO THE SCIENTIFIC STATUS OF ECONOMICS AND ECONOMETRIC METHODOLOGY

ADNAN VELIC

Junior Sophister

In the minds of many, to label any discipline as a 'science' affords it a certain level of legitimacy and authority. The question of whether economics can be classified as a science is a continuing source of contention within the field. In this paper, Adnan Velic discusses the criteria necessary for a subject to achieve scientific status, comparing the criticisms and counter arguments of conferring the standing of 'true science' on economics. Despite the contribution of econometrics, he acknowledges that critical shortcomings remain. The paper concludes by offering suggestions of how future practice might be improved upon.

Introduction

It is often stated that the Victorian historian, Thomas Carlyle, gave economics the derogatory nickname 'dismal science' in response to the late 18th century economic theory put forward by Thomas Robert Malthus. Malthus pessimistically forecasted that as population growth inevitably outstripped food supply, starvation would be the outcome (O'Hagan, 2000). Although he considered his principle of population as an economic theory of the past and present state of humanity, as well as a prognosis of the future, time did put it to the test. Irish experience, even during the Great Famine, refuted Malthus's theory as did the unexpected sensational advances in the efficiency of food production in the 20th century.

Since the time of Malthus, economics has made great progress in being recognised as a real science, yet it is still regarded by many critics as nothing more than a pseudo science. The economic analyst, Henry Phelps Brown, went as far as saying that what is wrong with modern economics is that its assumptions about human behaviour are randomly 'plucked from the air', thus rendering theories about real world economic phenomena non-scientific in nature (Brown,

1972).

This essay first seeks to examine the current scientific status of economics. It initially analyzes some of the properties that qualify a discipline to be labelled a science as well as reviewing some of the arguments of the critics who claim that economics does not deserve this title. We also scrutinise some of the counter arguments that seek to defend the field's standing as a true science. Afterwards, we look at the methodology of econometrics and how it has helped strengthen economics' claim to be a real science while assessing some of its weaknesses at the same time. Finally, we briefly consider some of the suggestions that have been made by those in the discipline on what could be done to help enhance the scientific status of economics.

The Scientific Status of Economics

The primary objective of a positive science is the construction of a theory or hypothesis that yields valid and meaningful predictions about phenomena not yet observed by a person (Friedman, 1953). One of the key features of a hard science is the scientific method that is used to test the theory in question and its predictions via the data obtained. This data lends empirical support to the theory where the test and its results are repeatable and demonstrable to others when the same conditions are present. Furthermore, scientists and philosophers, most notably Karl Popper, have asserted that no hypothesis can be considered scientific unless it is falsifiable. Falsifiability refers to the logical potentiality that a theory can be shown to be erroneous by an observation or a physical experiment. This principle follows from the fact that you can never demonstrate that anything is materially true but you can prove that some things are materially false. Popper utilises this fundamental asymmetry in developing his demarcation criterion (i.e. the distinction between science and nonscience) which states: science is that core of synthetic hypotheses about the actual world that can, at least theoretically, be falsified by empirical observations (Blaug, 1992).

Given the pre-requisites above and also the fact that scientific theories must be objective, it is necessary to distinguish between positive and normative economics. The development of all economic theories should be based on positive economics in order to render them scientific. These are 'what is' statements about the real world. However, normative economics deals with 'what ought to be' assertions which are essentially value-laden judgements based on a particular ethical position (Friedman, 1953). Although economic theory does not aim to make such value claims based on normative propositions, it is generally one of the reasons why economics is not recognised as being founded on

empirical observation and the scientific testing of hypotheses. Critics often use the argument that economics cannot achieve Popperian falsifiability due to the very nature of its study. Even though this is true in certain instances, many analysts contend that the problem is over-emphasised and that, more often than not, it is the practitioners who create the problem rather than the problem being innate to the nature of the subject. Instead of endeavouring to refute testable predictions of economic theories, modern economists appear content in illustrating that the real world accords with their forecasts, thus replacing falsification, which is difficult, with verification, which is simple (Blaug, 1992). Mark Blaug precisely depicts the problem with mainstream neoclassical economists:

> 'They preach the importance of submitting theories to empirical tests but they rarely live up to their declared methodological canons. Analytical elegance, economy of theoretical means, and the widest possible scope by ever more heroic simplification have been too often prized above predictability and significance for policy purposes. The working philosophy of science of modern economics may indeed be characterised as innocuous falsificationism' (ibid: 243).

According to Blaug, the problem now, is to entice economists to take falsificationism seriously.

Tests of theories in the physical sciences require that all conditions remain constant except for the experimental variable. Unfortunately we can rarely, if ever test certain predictions of theories in the social sciences such as economics by experiments explicitly designed to eradicate the most significant disturbing effects. In other words, it is almost impossible to conduct perfectly 'controlled experiments' in economics where all the relevant exogenous variables are taken into account and are kept constant in order to ensure a ceteris paribus effect is analysed between the two variables of interest (Friedman, 1953). Frequently, some of the factors are unobservable and cannot be taken into account. Thus, the ceteris paribus assumption, it is argued, is an over-simplification of reality. Critics maintain that the inability to administer controlled experiments in economics renders it a nonscience. Despite this contention, the counter-argument many put forward is that no experiment can be completely controlled and that every experience is partly controlled in the sense that only some disturbing effects are held relatively constant during its course e.g. experiments in astronomy in the physical sciences. The difference between controlled and uncontrolled experiments is thus one of degree (ibid). Therefore the inability to perform these so-called controlled experiments does not constitute

a major difference between economics and the physical sciences and does not render the theories any less scientific than those of the physical sciences.

In its attempt to be recognised as a real hard science, economics faces one major obstacle. As Lionel Robbins states, economics is 'the science that studies human behaviour' (Robbins, 1945). Human beings are not tadpoles that can be brought to a laboratory or cells that can be viewed under a microscope. It is impossible to get inside a human being or to perform repeatable experiments on human behaviour (Heywood, 2000). Thus the data that we can obtain about human conduct is limited and at times superficial. Such data is non-experimental and is often called observational data to elucidate the fact that the researcher is a passive collector of the information (Wooldridge, 2006). Given that many economic theories are based on, or somewhat related to human behaviour, we currently have no completely reliable means of testing economic theories in the absence of exact data (Heywood, 2000). Due to the fact that there is limited empirical evidence on human behaviour that can be used to refute or support economic theory, economics cannot just yet claim to be a real science.

An Appraisal of Econometric Methodology and its Interaction with Economics

Econometrics can be defined as the application of statistical and mathematical techniques to the analysis of economic data, with the aim of lending empirical significance to economic theories and verifying or refuting them (Maddala, 2001). The principal task of an econometrician is the development of relationships between different economic variables in mathematical form. This essentially provides us with a simplified model of the complex real-world process. After an economic theory has been transformed into an econometric model, the model is then tested with observed data (via t and F tests -'inference') and, if verified, is used for prediction and policy analysis. These are the two other main objectives of econometrics (Maddala, 2001).

For predicting the ramifications of changes, forecasting likely future outcomes and controlling variables to achieve targets, econometric models have a crucial function in modern economics and significantly improve the discipline's claim to be a real science (Hendry 1980). Despite these claimed improvements in the discipline, several analysts have characterised the methodology of econometrics as:

> 'an attempt to compensate for the glaring weakness of the data base available to us by the widest possible use of more and more

> sophisticated statistical techniques' (Leontief, 1971: 2).

Although this view is perhaps a bit cynical and outdated, it is true that in certain areas of economics distinct econometric studies can reach opposing conclusions. In addition, there are no adequate methods for determining which conclusion is valid given the notoriously unreliable data available. Subsequently, conflicting economic theories may continue to coexist for long periods. Such results are far from scientific. As such, at this point, it may be beneficial to examine some of the limitations of econometrics in assisting economics on its route to becoming a science.

One reason why econometrics and therefore economics cannot be declared to be being scientific is because of the conditional nature of economic relations. In general these relationships appear to be time specific and hence fragile and unstable 'when exposed to the light of day' (Johnston, 1991). This renders economic theories supported by econometric results non-scientific when compared to the requirements that have to be satisfied by theories in the natural sciences. An example of this was the failure of the original Phillips curve after the oil shock of 1973. This point was impressively illustrated by Lord Robbins whilst scrutinising the disparities between economics and the natural sciences:

> 'The influence of the Reformation made no change in the forces of gravity. But it certainly must have changed the demand for fish on Fridays' (Johnston, 1991: 53).

Studies conducted in the natural sciences yield unconditional predictions and are beyond the domain of being effected by human actions while those in economics and econometrics are not. However, we cannot use this argument as concrete evidence of the non-scientific nature of econometrics; some predictions of theories in the natural sciences are conditional too. Hence we must make a further distinction between strong and weak conditional predictions. Typically, scientific predictions are strong which means that under specific conditions an event will occur with certainty. In contrast, economic forecasts are weak meaning that an event will follow if there is no disturbance i.e. some of the conditions for the event to follow have been detected (McGrath, 2002). The difference is due to the fact that human behaviour and the forces present in the field of economics are both unpredictable and erratic in nature across time. Consequently, we surely cannot expect econometricians to develop 'super equations' that take into account all the relevant factors and conditions (Johnston, 1991). Hence the nature of the study of economics renders the methodology of econometrics non-scientific and economics a nonscience when

compared to the natural sciences.

The scientific method, when testing theories in the natural sciences, requires that the test be unbiased. However, complete objectivity is very difficult to achieve in econometrics because during the specification aspect of the econometric work, it is often up to the econometrician to select the functional form of the regression model. Hence, deception is easy as econometricians can exploit this opportunity and choose the functional form of the model that corroborates their own economic theory (Hendry, 1980). There is no doubt according to some critics that such regression analysis and testing of economic theories relies on cookbook econometrics where a hypothesis is expressed in terms of an equation, a variety of forms is estimated for that equation, the best fit is selected while the rest are rejected and then the theoretical argument is modified to rationalise the hypothesis that is being tested (Ward 1972).

Econometric methodology provides us with probabilistic results due to the need for a stochastic error term (to account for the variables that we cannot include in our model which influence the endogenous variable of interest). However, econometrics relies too much on experimentally unobservable human behaviour and on the assumption that unforeseen events or disturbances follow statistical distributions (Hendry, 1980). Hence, due to the greater uncertainty that is attached to econometric results, econometrics cannot be recognised as an authentic definite science.

Enhancing the Scientific Status of Economics

Given some of the problems associated with econometrics, there have been those who have suggested that these are justifications for abandoning it altogether (Blaug, 1992). However such action would leave economics with practically no means of choosing, from an abundance of possible theories, the one theory that best explains the economic event in question. Although there do exist other techniques for testing economic theories, such as ethnographic methods etc, the requests of state economic policy makers will always bring us back to the methodology of econometrics (Blaug, 1992). Rather than dispose of something that has done more good than harm in advancing the scientific status of economics, numerous suggestions for strengthening the claim of economics and econometrics together as a hard science have been made by Thomas Mayer. Some of these recommendations are outlined below.

Mayer first suggests that we place greater emphasis on data collection. Second, he asserts that econometric results should not be considered as evidence from a ‘crucial experiment’ which is not to be repeated but rather that applied

econometrics should aim to replicate past results using different data sets (Blaug, 1992). As Hendry states, the three golden rules of econometrics are 'test, test and test' (Hendry, 1980). This would do much to resolve the problem of contradictory hypotheses coexisting by relying on many pieces of evidence being pulled together from many periodic tests rather than falling back on the results of a single crucial experiment. Third, we should try and eradicate data mining by requesting authors to present all the regressions they ran and not just the specific regression that happened to support their economic theory (Mayor, 1980). In addition to these suggestions, Johnston recommends that in order to improve the quality of econometric work in the short-run, a greater balance needs to be achieved between theoretical and applied work in econometrics (Johnston, 1991). According to him, more empirical work with the objective of verifying economic theories would go a long way to enhancing the scientific status of economics and econometrics as a measurement tool.

Conclusion

As we have seen throughout the course of this essay economics cannot yet be regarded legitimately as a real science although econometrics has contributed vastly to this cause. Economics faces a variety of difficulties in achieving this objective due to the substandard practices of those in the discipline who often fail to exercise the methodology they preach and also as a consequence of the problems that arise in the field which are innate to the nature of the subject. The artificial obstacles constructed by the economists and econometricians themselves can indeed be eliminated by imposing more stringent measures in terms of methodology in order to augment the scientific status of the field. Too often our researchers who are protective of their own theories fail to ask the scientific questions such as; does there exist a different model that fits the data well and that is better at explaining the economic phenomena in question? Instead they pose the same question that a juggler's spectators would ask; have virtuosity and skill been exhibited? (Summers, 1991). However, the natural complications that arise in economics are more difficult to overcome even with the aid of econometrics but despite this, theory based on empirical evidence is surely the discipline's best defence against any criticisms seeking to diminish its scientific status.

Bibliography

Blaug, M. 1992. *The Methodology of Economics: or How Economists Explain*. Cambridge: Cambridge University Press.

Friedman, M.J. 1953. *Essays in Positive Economics*. Chicago: University of Chicago Press.

Hendry, D.F. 1980. 'Econometrics-Alchemy or Science?'. *Economica* 47:188:387-406.

Heywood, A. 2000. *Key Concepts in Politics*. New York: Palgrave Macmillan.

Johnston, J. 1991. 'Econometrics Retrospect and Prospect'. *Economic Journal* 101:404:51-56.

Leontief, W. 1971. 'Theoretical Assumptions and Nonobserved Facts'. *American Economic Review* 61:1:1-7.

Maddala, G.S. (ed.) 2001. *Introduction to Econometrics*. Sussex: John Wiley and Sons Limited.

Mayer, T. 1980. 'Economics as a hard science: realistic goal or wishful thinking'. *Economic Inquiry* 18:2:165-178.

McGrath, G. (ed.) 2002. 'Econometrics: Pro-Active Science, or PR Stunt?'. *Student Economic Review* 2002, Dublin.

O'Hagan, J.W. (ed.) 2000. *The Economy of Ireland: policy and performance of a European region*. Dublin: Gill and Macmillan.

Phelps Brown, E.H. 1972. 'The underdevelopment of economics'. *Economic Journal* 82:325:1-10.

Robbins, L. 1945. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan and Co., Limited.

Summers, L.H. 1991. 'The Scientific Illusion in Empirical Macroeconomics'. Scand. *Journal of Economics* 93:2:129-148.

Ward, B. 1972. *What's Wrong With Economics?* London: Macmillan.

Wooldridge, J. (ed.) 2006. *Introductory Econometrics: A Modern Approach*. Mason: Thomson South-Western.

THE NEW ECONOMIC HISTORY: A DISCUSSION OF CLIOMETRICS

SIMON MEE

Junior Sophister

To plan for our future, it is necessary to learn from our past. However, since so much of history is essentially based on perceptual bias, how are we most effectively to learn from this past? In this paper, Simon Mee explores the evolution of cliometrics, 'the new economic history'. He examines how it came into existence and how it differs from the traditional school of economic history. Through an analysis of the 'slavery debates' he goes on to describe the key methods of this field of study. Finally, the essay questions the merits and contributions of cliometrics to economic history and concludes that by combining both qualitative and quantitative perspectives we can learn the most from our economic past.

Introduction

'If cliometricians were asked to write a history of the crucifixion, they would begin by counting the nails'[1]

The study of economic history can be traced back to Adam Smith, who balanced his theory and reasoning with frequent reference to statistical and historical illustration. Economic history has in the past acted as an antidote to economists' reliance on abstract theory. It is a combination of two disciplines that has experienced a curious evolution since the time of classical economics. The introduction of cliometrics represented a significant break from the past. In the period after the Second World War, a sizeable shift occurred in the methods and techniques applied in approaching economic history. Aided by the arrival of the computer, these new approaches were the product of a younger generation of economic historians who had been shaped and influenced under the pressure of

[1] Eugene Genovese, quoted in (Fogel, 2003: 22)

the particular historical situation in which they grew up. Famous cliometricians such as Robert W. Fogel, Douglass North, Davis and Alfred H. Conrad were all born within eight years of each other (Redlich, 1965).

The work of this generation soon came to be known as the 'new economic history', often referred to as 'econometric history' or 'cliometrics', a reference to the muse of history, Clio.[2] This essay will attempt to distinguish cliometrics from the traditional school of economic history, and look at its emergence from the late fifties onward. Through a critical discussion of the 'slavery debates' the essay will then go on to examine the key features behind cliometrics, including the growing emphasis upon quantitative constructs and the systematic use of economic analysis in applying these constructs. The essay will conclude with an evaluation of cliometrics' contribution to economic history.

Cliometrics and the Traditional School of Economic History

Cliometrics is the systematic application of the behavioural models of the social sciences and their related mathematical and statistical methods to the study of history (Fogel, 2003). Quantitative techniques are used to make interpretations and reconstructions of the past. Whereas much of the traditional economic historian's work was limited to the location and simple classification of existing data, the cliometrician actively 'reconstructs measurements … no longer extant' (Redlich, 1965: 482). They began using methods to combine existing data with new measurements in order to gain better insight into the topic at hand. Most economic history up to this point had remained primarily qualitative, with numerical information used largely as illustration. The training of historians did not tend to direct them towards the discovery of quantitative records. Indeed these records were most likely to be found in government offices, business firms and savings institutions. The interpretation of such documents often required certain technical skills that the ordinary historian simply did not possess (Conrad and Mayer, 1965). Cliometricians on the other hand, often came from a background in economics, and set out to reconstruct American economic history on a quantitative basis.

Thus cliometrics was deemed a controversial approach by the academic community; many historians refused to accept it as good history. The sheer emphasis on quantitative measurements and techniques tended to confound historians from the older generation. Since it was solely based on concepts and data, it seemed to disregard one of the most important sources favoured by

[2] In Greek mythology, Clio is the muse of history.

historians; primary sources (Engerman, 1977). The overriding concern with economic determinants led to a neglect of the various moral judgments which until then had featured prominently in American historiography. In a broader sense, traditional economic historians had primarily dealt with the development of economic institutions, with a secondary focus on the processes taking place therein. Cliometrics however tended to deal primarily and directly with the economic processes, while neglecting economic institutions.

One of the factors underlying the controversy surrounding cliometricians were the substantive conclusions they had established, which in effect challenged the well-established propositions of traditional historiography. Had cliometrics produced the same conclusions, there may well not have been so much debate and notoriety surrounding it. Some traditional economic historians reacted as if they were being collectively accused of blunders in their work. As Fogel remarks, cliometricians 'refused to be bound by the established rules of engagement, and they blithely crossed ideological wires in a manner that perplexed and exasperated traditional historians….' (Fogel, 2003: 19).

The Slavery Debates

Until the 1950s, discussion of the economic consequences of slavery was dominated by the work of Ulrich B. Phillips. Phillips' influential book, *American Negro Slavery*, was first published in 1918 and went on to become the accepted canon of knowledge on the nature of the American slave system. The 'Phillips school' maintained that slavery was an inefficient system that stifled the growth of the South in the Antebellum Period, and that by the time of the Civil War the institution was moribund (Poulson, 1981). He purported that the main purpose of plantation slavery was not economic, but social; slavery was kept in existence because of the transitory resolve of a class long accustomed to its peculiar social institutions (Fogel, 2003).

This interpretation of American slavery went unchallenged until 1956, when the publication of Kenneth M. Stampp's *The Peculiar Institution* attacked the inadequacies of Phillips' argument. Stampp denounced Phillips' emphasis on the 'benevolence' of slaveholders and instead portrayed the plantation system as merciless and exploitative (Smith, 1998). Interestingly, Stampp went on to conclude that slavery was actually an efficient economic system. This was an assertion which generally passed unnoticed at the time, but it nonetheless caught the attention of some up-and-coming cliometricians. While the older generation of historians received the work with scepticism, it was enthusiastically embraced by the young revisionists. It was this research, along with the seminal work of

Alfred H. Conrad and John R. Meyer that helped to shape the debate which was to follow over the next three decades (Conrad and Meyer, 1958).

Conrad and Meyer's work gave cliometrics its first formulated expression of the new approach; quantitative method could be used to verify a qualitative hypothesis. From then on, a steady flow of papers using this new approach appeared with increasing rapidity. The most controversial of these contributions to the literature on the southern economy were the studies of Robert W. Fogel and Stanley L. Engerman. In 1974, *Time On The Cross* was published, resulting in national fame and notoriety for the authors. The book was as controversial for its reliance on certain statistical data as it was for its findings. Within *Time On the Cross*, three central topics were addressed: the profitability and economic viability of slavery, the rate of southern economic growth between 1840 and 1880 along with the factors which influenced it, and the relative productivity of slave and free agriculture (Fogel and Engerman, 1974). Through quantitative methods, such as regression analysis, they found that the productivity of southern farms exceeded that of northern farms by roughly thirty-five percent, and that the majority of this difference could be explained by the greater efficiency of the slave plantations. Between 1840 and 1860, per capita income actually increased more rapidly in the South than in the rest of the nation. Furthermore, the Negro slave produced as much output in thirty-five minutes as a free farmer did in a full hour (ibid). They argued that this was due to the intensive utilization and specialization of the slave labour force; under the plantation system the slave labour force was highly coordinated.

Fogel and Engerman's book attacked Phillips, Genovese and other historians in their interpretation of the South as a pre-capitalist society. *Time On The Cross* depicted the slaveholders as behaving like rational businessmen, who ordered their plantation work regimens along the lines of northern factories. In other words, the plantations were treated as capitalist enterprises where profit was the underlying motive, not social reasoning. One example of this they argued, could be seen through slave prices: ten year old slaves were cheaper than twenty-six year olds, who were priced twice as high. This was because the latter cost the planter less money to rear and because the older slave procured a higher return.

The authors of *Time On The Cross* refused to let moral judgment interfere with their argument. For them, it lay outside the economic realm. Some historians however, have criticized Fogel and Engerman's application of economic models to slave society. Was it appropriate to apply the concepts of economic exploitation and economic efficiency to a mode of production based upon involuntary servitude? Was it right for the qualitative to be measured by quantitative methods? These questions directed the economic historian towards

an examination of the features which lay behind cliometrics: the growing emphasis upon quantitative constructs and the systematic use of economic analysis in applying these constructs. As they discovered, cliometrics is far from a perfect approach.

Counting the Nails: Cliometrics and Methodology

With regard to the use of quantification, the 'new economic history' is a somewhat misleading name. It has been argued that there is nothing 'new' about the approach. In his Boston address of 1963, Fogel asserted that the effort 'to rediscover and present numerical information relative to historical processes' was not a recent one (Redlich, 1965: 482). The foundations of the quantitative methods go back to the thirties and forties, with men such as John Clapham and Walt W. Rostow (Davis, 1966). However these were isolated attempts, and though data was available for some time, the techniques required to analyse and interpret it systematically had not been perfected until after the Second World War (Fogel and Engerman, 1974). Fogel later went on to say that the innovative aspect of the cliometrician's work was the actual approach to measurement and theory; the underlying process through which they applied this economic data. Nonetheless, the emphasis on quantitative methods has its disadvantages. Figures and numbers themselves do not represent the processes of the cliometrician at work. Figures are quantitative symbols which stand for something: the *result* of a process. By using these figures in time series, or through regression analysis, an impression is incorrectly given that they represent the process, whereas in reality they merely acted as 'yardsticks' (Redlich, 1965).

In addition, restricting the emphasis to quantitative constructs imposed something of a limit on any possible analysis or interpretation which could occur. It was often noted that many cliometricians were not as willing as Fogel to 'get their hands dirty' in terms of reconstructing data (Davis, 1966). There tended to be a concentration of work based on the public sector, due to the fact that such data was already available. This concentration therefore tended to mirror not the importance, but the relative availability of data (ibid). As a result, cliometricians tended to draw strong criticism from traditional historians, who accused them of 'easy work' (Engerman, 1977: 79).

The possible restrictions posed by the lack of reliable data are quite evident in Conrad and Meyer's influential essay *The Economics of Slavery in the Ante-Bellum South* (1958). One critic found that in the first section of their essay, Conrad and Meyer resorted to heaping dozens of additional assumptions on top of the basic one in order to make the data suitable. They were in effect twisting

their model, and ultimately their conclusion, to suit their data. Thus it is not difficult to see why cliometrics received its share of criticism. The choice of model used by the cliometrician was to be a constant source of debate between those economic historians trained as economists and those trained as historians. It was a crucial assumption that fundamentally affected the outcome - and possible revision - of economic history more than any new detailed examination of data, for it is the choice of model through which the data is analysed.

Yet the shift in the extent and nature of quantitative measurement was to be seen with the publication of *The Economic Growth of the United States, 1790-1860* by Douglass North (1961). It was the nature of the subject matter which led this book to be much debated. North stated in the preface of his book that due to the preoccupation with description and institutional change, there had yet to be a comprehensive and integrated analysis of the United States. While it previously had been custom to provide separate treatments of various sectors within the economy, North attempted to capture economic growth as a whole. His work represented a fundamental expansion in the remit of quantitative analysis and in turn inspired many young economic historians to push the boundaries of cliometrics.

However, it was the use of 'counterfactual hypotheses' for which cliometricians were notorious. Essentially, the counterfactual hypothesis attempted to establish and measure what *could* have happened in order to understand what *did* happen. An example of this was when Robert Fogel tested a widely accepted thesis by asking whether the railroads were really the central feature in American development (Fogel, 1964). Controversially, he designed a model of the nineteenth century United States without railroads and found that America's development would not have changed much since alternative methods of transportation would have taken over.

This finding caused a furore among the academic establishment. Redlich accused Fogel of attempting 'quasi-history', in that counterfactuals were fundamentally alien to economic history (Redlich, 1965). However, Fogel countered with the argument that the traditional economic historian abounds in disguised counterfactual assertions, citing the example of traditionalist essays which argue that slavery retarded the development of the South (Fogel, 1971). In Fogel's view, the difference between the old and new approach 'is not the frequency with which one encounters counterfactual propositions, but the extent to which such propositions are made explicit' (ibid: 10). The counterfactual brought difficulties into the analysis in terms of deciding where it was necessary to draw the line once some changes were introduced, and of defining the time period over which the underlying assumptions seemed acceptable.

Cliometrics and Theory

There were rapid advancements made in the field of economic theory during the post-war period. Advances in the theory of economic development allowed cliometricians to measure the growth of nations more accurately. National income had only recently become an operational construct, due to the influential work of Wesley Mitchell and Simon Kuznets at the National Bureau of Economic Research. It was only in 1946 that estimates of national income back as far as 1869 were made available (Engerman, 1977).

Through increasing emphasis on theory - particularly neoclassical economic theory - cliometricians were able to circumvent the possible problem of poor data. Theory allowed the cliometrician to analyse data in the context of a given framework. This introduction of neoclassical theory into economic history led to a heated debate: the theoretical world in which market forces and totally rational human beings operate without friction is a long distance from the actual world in which societal rules and customs, as well as complex human motivations, interfere heavily with economic phenomena. Indeed, to subject the past to utility maximisation was in effect to argue in favour of understanding the economic past as a totally observable environment.

Conclusion

It is clear that cliometrics is not without its faults. Some questionable works have appeared due to the tendency of some cliometricians to assume that it is only necessary to apply economic theory with quantitative techniques and call it history. While it is true that some problems can be traced to poor technical practice, it must be noted that bad work does occur in all scholarly research. It is often forgotten that the historical past is not a plain truth; rather it is a construct and interpretation of the person who chooses to write about it. One can ask if history - economic or otherwise - could ever be objective. However, to imply that quantitative techniques and data could offer the objective purity that the reader desires would be misleading. As has already been seen, the quality of the data and the econometric model will always distort the conclusion drawn, allowing the cliometrician to make what he will of the past, just as any other historian can. History itself is far too intricate, far too complicated a process to be analysed through numbers and quantitative techniques alone. After all, there was more to the crucifixion than the number of nails.

Nevertheless, one cannot deny the clear advantages that cliometrics has offered; it has expanded the boundaries and techniques of what economic history

can achieve. It is true that, over the years, cliometrics has evolved in its approach with the proper use of quantitative methodology, and it has been this maturity which has allowed cliometrics to establish itself as a proper school. It was a reaction to the traditional economic history that came before it, and in turn has gone on to shape historical research and writing to this day. The two schools have travelled a long way since their acrimonious struggle of the sixties and seventies; passions, egos and tempers have long since calmed. Critics at the time argued that cliometrics was too limited, too rigid an approach to allow for proper historical debate. They said its emphasis on quantitative methods led to an excessive narrowing of the question. With hindsight however, specialized knowledge can help add pieces of history together to provide answers to broader questions. After all, it was Adam Smith who pointed out that specialization and division of labour are not without benefits when working towards a common goal. Indeed, it is hard to ignore the substantial interdependence that has since emerged between the two approaches over the years. By maintaining a delicate balance between the two schools, the economic historian can have the freedom of qualitative judgment while using quantitative methodology to support his argument.

Bibliography

Conrad, A.H. and Meyer, J.R. 1958. 'The Economics of Slavery in the Ante-Bellum South'. *The Journal of Political Economy* 66:2:95-130.

Conrad, A.H. and Meyer, J.R. 1965. *Studies in Econometric History*. London: Chapman & Hall.

Davis, L. 1966. 'The New Economic History II. Professor Fogel and the New Economic History'. *The Economic History Review* 19:3:657-663.

Engerman, S.L. 1977. 'Recent Developments in American Economic History'. *Social Science History* 2:1:72-89.

Fogel, R.W. 1964. *Railroads and American Economic Growth: Essays in Econometric History*. Baltimore: John Hopkins Press.

Fogel, R.W. 1971. *The Reinterpretation of American Economic History*. New York: Harper & Row.

Fogel, R.W. 2003. *The Slavery Debates, 1952-1990*. Louisiana: Louisiana State University Press.

Fogel, R.W. and Engerman, S.L. 1974. *Time On The Cross*. London: Norton.

North, D. 1961. *The Economic Growth of the United States, 1790-1860*. London: Norton.

Poulson, B.W. 1981. *Economic History of the United States*. New York: Macmillan.

Redlich, F. 1965. '"New" and Traditional Approaches to Economic History and Their Interdependence'. *The Journal of Economic Science* 25:4:480-495.

Smith, M.M. 1998. *Debating Slavery. Economy and Society in the Antebellum American South*. Cambridge: Cambridge University Press.

FINDING NEW LINKS – FISHER'S EQUATION OF EXCHANGE & NEWCOMB'S EQUATION OF SOCIETARY CIRCULATION

MICHAEL CURRAN

Senior Sophister

The Quantity Theory of Money, though admittedly controversial, is widely accepted as one of the cornerstones of modern economic thought. Michael Curran traces the development of this prominent theory by examining the works of Simon Newcomb and Irving Fisher. In a fitting tribute to two of our great economic theorists, he shows how Fisher's equation of exchange has built on Newcomb's equation of societary circulation, ultimately propelling the QTM to the influential position it holds today.

Introduction

The Quantity Theory of Money (hereafter QTM) has appeared sporadically: 'dating back at least to the mid-sixteenth century Spanish Scholastic writers of the Salamanaca School' (Humphrey, 1997: 71). In 1522, Nicolaus Copernicus first explained the principles on which to base a currency. A more complex version of Copernicus' simple QTM was provided by the French social philosopher Jean Bodin in 1568. Classical economists stressed the application of the ordinary theory of supply and demand to money in the QTM; given real demand for money, an increase in the nominal supply of money will lower the value of goods that each unit of money can purchase. Irving Fisher emphasised the inverse relationship between this value, i.e. the purchasing power of money, and the quantity of money.

This paper investigates the familiar equation of exchange arising from Fisher's 1911 *The Purchasing Power of Money* (PPM). The equation of societary circulation, anticipated by Newcomb in *Principles of Political Economy* (1885), is also discussed.

Irving Fisher

Paradoxically popular both at Yale and Chicago today, the theorist-reformer polymath Irving Fisher (1867–1947) chose economics as a career path over mathematics, like Milton Friedman. Fisher's was the first Yale PhD in pure economics, albeit it was awarded by the faculty of mathematics. His tutors, or advisors, included Willard Gibbs; physicist, and William Graham Sumner; economist.

Irving Fisher's interest in solving monetary problems led to his developments in econometrics and statistics.[1] Fisher was the 'greatest expert of all time on index numbers' (Tobin, 1987: 369) By collecting data globally, Fisher's Index Number Institute produced price indices from 1923 to 1936. His dedication to increasing the availability of statistical data foreshadowed the emergence of the Penn World Tables in the 1980s.

Robbins (1998) claims that Fisher's most important work was carried out before the First World War. Despite the originality and innovative nature of Fisher's work in the 1920s, such as his theory on the affect of the rate of change of prices on output and employment, it failed to attract contemporary attention. This was possibly due to the fact that he published many of these later articles in obscure journals, when highly technical economics had not yet become mainstream (Blaug, 1992).

The Purchasing Power of Money

The purpose of Fisher's PPM was to explain the five determinants of the purchasing power of money and then apply the theory to the study of historical changes in this purchasing power.[2] It is a restatement of the QTM, which he a sserted as dogma. He discredited economists who questioned the validity of the QTM, e.g. Laughlin, head professor of political economy at Chicago. At the 17th annual meeting of the AEA in Chicago, December 1904, Fisher gave credit to Newcomb: 'The purchasing power of money is dependent on what Professor Newcomb calls 'the equation of societary circulation'' (Fisher, 1905: 122). He fervently promoted the QTM:

---

[1] Some of Fisher's analyses contain probably the first use of distributed lags in economics.

[2] Fisher's *PPM* was dedicated: 'To the memory of Simon Newcomb … Great Scientist, Inspiring Friend, Pioneer in the Study of Societary Circulation'.

> 'The real problem to be settled is not the quantity theory, which, in the sense that I have described, seems to me, should be accepted at the outset but to proceed further and discover the other quantities in this equation' (ibid: 124).

Friedman dubbed the transactions version of the quantity equation as the most famous and claims that Newcomb 'formulated' it and Fisher 'popularised' it (Friedman, 1987). The equation:

$$MV = PT \quad [1]$$

relates the circular flow of money, over a specified period of time in a given economy, to the circular flow of goods. Money is a stock, *M*, the dimensions of which are dollars. The dimension of *V* is the number of turnovers per unit time, so *MV* is a flow. The real value of money-using transactions, *T*, is a flow since its units are the number of unit quantities per unit time, while P represents the price level. The left-hand side of the equation can be further expanded into payments using either currency, *MV* or deposits, *M'V'*. The choice of this subdivision depends on whether or not deposits (in addition to currency) are included and on the availability of data on *M'V'*. Fisher identifies five direct influences on the general price-level *P*: *M*, *M'*, *V*, *V'* and *T*. The price-level, *P*, is the only dependent variable in the equation of exchange:

$$MV + M'V' = PT \quad [2]$$

To derive the QTM from the equation of exchange, Fisher argued the following: T is exogenous; velocities (*V* and *V'*) are functions of institutions and habits and are independent of the other variables; the relation between *M* and *M'* is fairly stable and is also independent of the other variables; and finally, that exogenous changes in currency supply are the principal source of shock. As a result of these conditions, changes in the currency supply move *P* proportionately. Fisher's belief in the slowness of institutional change helped to support his notion that the velocities are stable.

Fisher explained his monetary transmission mechanism in *The Purchasing Power of Money* (1911). If the ratio *T/V* is fixed, raising the money supply will cause the supply of money to exceed demand. People will dispose of their excess supplies of money by increasing their demand for every good. The result is a price rise for all goods, and thus *P* rises. This process continues until the demand and supply of money are equalised. Fisher's explanation rationalised

the left-to-right causation whereby prices rise proportionately to meet the higher money supply; this is Fisher's identification of the QTM (Fisher, 1911b). Interestingly, it is the interaction of markets, similar to Keynes' approach, through which an excess supply of money leads to excess demand for commodities.

Two particularly strong assumptions underlie the equation of exchange: first, that the economy is at the natural rate of unemployment; secondly, that *V* is stable – predictable, but not necessarily constant. The implication of these two assumptions, combined with the equation of exchange, is that 'inflation is always and everywhere a monetary phenomenon' (Friedman and Schwarz, 1963: 592). The policy implications of this analysis, while predating Friedman, may suggest monetary targeting. Firstly, the authorities ought to control the rate of change of money supply. Secondly, with a stable $V = PT/M$, PT should be controlled via M. However, there are two significant problems with this approach. One concerns the stability of *V*; if *V* is unstable, it becomes difficult to control PT via M. The other pertains to problems of reverse causation with M being endogenous, or, even simultaneity – two way causation.

By studying price data, Fisher knew that in reality prices are imperfectly flexible, meaning that *T* would adjust, absorbing some of the changes on the left-hand side of the equation of exchange. Fisher also discusses non-neutralities of real interest rate movements on the real economy; however, these transitional effects are not emphasised, so as not to detract from his conclusion. Schumpeter (1951) believed that 'the scholar was misled by the crusader'; to promote the compensated dollar, Fisher needed a simple plan for stabilizing purchasing power:

> 'Had Fisher pulled these strands together into a coherent theory, he could have been an American Keynes. Indeed the 'neoclassical synthesis' would not have had to wait until after World War II. Fisher would have done it all himself' (Tobin 1987: 376).

While Fisher's work is a stepping stone from money to employment, he did not fully integrate his theory of M with that of P and distributive shocks. In order to advocate the QTM, he disguised his recognition of the fact that fluctuations in money temporarily influence V; hence his hypothesis was that *V* is an institutional constant. The specific factors determining P were relegated to the background. He did not ignore the volatility of *V*; however, P varies directly with M only if *V* and *T* are constant and varies in the same proportion only in the 'ultimate' and 'normal' sense.

Much of Fisher's *The Purchasing Power of Money* is devoted to a quest of aggregation, specifically that of making equation [1][3] compatible with:

$$MV = \Sigma p_i Q_i \qquad [3]$$

where p_i and Q_i are individual prices and quantities. Fisher recognized the heterogeneity of transactions. He wanted to find aggregate indexes *P* and *T* such that the two forms of the equation would be consistent. His pioneering work in this area led him to choose the Paasche price index in 1911, which is based on future consumption:

$$P_1 = \frac{\Sigma p_1 Q_1}{\Sigma p_0 Q_1} \qquad [4]$$

over the Laspeyres price index:

$$L_1 = \frac{\Sigma p_1 Q_0}{\Sigma p_0 Q_0} \qquad [5]$$

Challenged by Cambridge school economists, the popularity of Fisher's *The Purchasing Power of Money* waned soon after it was published. However, Trautwein supports the notion that recent decades have seen a revival of Fisher's ideas: 'Fisherian concepts … now permeate much of modern macroeconomics, finance and econometrics' (Trautwein, 2007: 609). The issues of index numbers Fisher was concerned with do not appear as important today according to Tobin (1987); similar to Kenneth Arrow's seminal Impossibility Theorem, in 1931 Frisch proved that there cannot exist a price index satisfying the axioms that Fisher believed were sensible. However, there is resurgence in this field; recent works display interest in this area, e.g. Somerville (2004).

[3] *T* is a measure of all transactions involving he offer of money: intermediate as well as final goods and services, old and new goods, financial assets as well as goods. Fisher's preference for a comprehensive measure and concept of *T* has not been developed to any significant extent by later writers.

Simon Newcomb

Simon Newcomb (1835–1909) was a world-leading astronomer[4], and a major American mathematical economist preceding Fisher. His interest in economics was first sparked in 1854 by reading Say's *Political Economy*, eventually lecturing in the subject at Harvard. In an attempt to explain why Newcomb's work didn't attract the attention it deserved, Fisher conjectured that:

> 'once a man's name becomes associated with a particular department of knowledge like astronomy, any attempts to contribute to other departments encounter a prejudice which it is difficult to overcome' (Fisher, 1909: 641).

Transforming Newcomb's Equation of Societary Circulation into Fisher's Equation of Exchange

The P in Newcomb's equation of societary circulation:[5]

$$VR = KP \qquad [6]$$

refers to the scale of prices, as distinct to particular prices, and hence anticipates Fisher's emphasis on this difference. Instead of Newcomb's 'rapidity' of circulation, R, apparently coined by Thornton, Fisher uses 'velocity', which suggests natural stability. Newcomb's total volume of currency, *V*, which is the sum of coins, paper money and deposits, was transformed by Fisher into *M* and *M'* with corresponding velocities, *V* and *V'*, respectively[6]. Emphasised by empirical estimates (Fisher, 1911a, 1911b, 1912, 1913), Fisher's measure of *T* is implicitly broader than Newcomb's *K*. He used bank clearings as a measure of *M'V'*; this measure included large volumes of transactions in financial assets. He mostly discusses changes in *T*, which refers to changes in output and employment (Fisher 1911a). Newcomb excluded borrowing and lending from his measure of the transactions as well as speculative transactions (Newcomb, 1885).

The national income interpretation is plausible for Newcomb; his identification of *K* sounds like national output, or income. However, the central theme of his book concerns the nature, determinants and consequences of

[4] Mentioned in H.G. Wells' *The Time Machine* (1895).
[5] Fisher's *MV = PT*.
[6] Bordo (1987) incorrectly refers to Newcomb's *V* as the total 'value' of currency.

production. As an illustration of Newcomb's national income orientation, if there is a one per cent fall in expenditure flow: '… one per centum of the industrial population would be thrown out of employment so long as the scale of prices remained the same as before' (Newcomb, 1885). In contrast, an income interpretation of Fisher's work is not easy to support. Fisher never incorporated national income data into the equation of exchange. The equation is still in transactions form in his Booms and Depressions (Fisher, 1932).

For a monetary transmission model, in particular to look at how changes in V affect P, Newcomb created an aggregate demand function:

$$D = \frac{NF}{P} \qquad [7]$$

where the flow of the currency, F, is defined as:

$$F = VR \qquad [8]$$

Substituting the expression for F into the aggregate demand function yields the equation:

$$D = \frac{NVR}{P} \qquad [9]$$

where D is the quantity of goods demanded and N is a fixed constant; hence D is directly proportional to V/P and does not depend on equiproportionate changes in both V and P (Newcomb, 1885). Thus, an increase in V initially leads to a rise in real demand, D, which forces a rise in P and eventually the initial expansion of V causes P to rise by the same proportion; hence, real demand, D, returns to its original level. Money is neutral and the QTM is satisfied since the proportionate change in V and P are the same, with DP responding to DV. V has no effect on the steady state values of the real variables K and R.

Although Newcomb promoted the QTM, he stated the necessary condition that the theory holds only if prices are flexible. If prices are sticky monetary non-neutralities occur as a result of K needing to adjust to bear some of the change in M. Newcomb noted that holding R and K constant, an exogenous rise in price would lead a compensatory fall in K. Despite this, he warned policy makers not to respond to these price increases to guarantee full employment by expanding the money supply, lest inflation would follow. Retrospection suggests an anticipated corollary of this – the short-run Philips curve.

Conclusion

Irving Fisher, the most cited economist in the world in the early twentieth century,[7] has contributed prodigiously to modern economics. Surprisingly his work has often been regurgitated without accolade.[8] The quantity equation, as espoused by Newcomb, owes its dominant position in popularity over all other quantity equations to Fisher. Its re-discovery has coloured the field of monetary economics. On the future of the debate concerning the QTM, Friedman's conclusion seems apt:

> 'One thing is certain: the quantity theory of money will continue to generate agreement, controversy, repudiation, and scientific analysis, and will continue to play a role in government policy during the next century as it has for the past three' (Friedman, 1987: 19).

Bibliography

Blaug, M. 1992. *Pioneers in Economics Vol. 41.* Aldershot: E. Elgar.

Bordo, M.D. 1987. 'Equation of Exchange' in *The New Palgrave: A Dictionary of Economics* 2:175–177.

Dimand, R.W. and Geanakoplos, J. (eds) 2005. *Celebrating Irving Fisher: The Legacy of a Great Economist.* Malden, MA, Oxford, and Victoria: Blackwell Publishing.

Fisher, I. 1905. 'Discussion of the theory of money, 17th Annual Meeting of the AEA December 28–30 1904' reported in *Publications of the AEA 3rd Term* Vol:VI:1:122–5.

Fisher, I. 1909. 'Obituary: Simon Newcomb' *Economic Journal* 19:76:641–644.

[7] See Dimand and Geanakoplos (2005) on different studies of Fisher's rankings in citations.
[8] The chapter on capital markets in Gravelle and Rees (2004) offers an excellent exposition of an area of Fisher's theory including the 'Fisher separation theorem'; alas, they have severed his name tag from the entire chapter.

Fisher, I. 1996. [1911a]. 'The Purchasing Power of Money: Its Determination and Relation to Credit Interest and Crisis' in W.J. Barber (ed.) *The Works of Irving Fisher*. London: Pickering and Chatto.

Fisher, I. 1911b. 'The Equation of Exchange, 1896-1910'. *American Economic Review* 1:2:296-305.

Fisher, I. 1912. '"The Equation of Exchange for 1911", and Forecast'. *American Economic Review* 2:2:302-19.

Fisher, I. 1913. '"The Equation of Exchange" for 1912, and Forecast'. *American Economic Review* 3:2:341-5.

Fisher, I. 1996. [1932]. 'Booms and Depressions' in W.J. Barber (ed) *The Works of Irving Fisher.* London: Pickering and Chatto.

Friedman, M. 1987. 'Quantity Theory of Money' in *The New Palgrave: A Dictionary of Economics* 2:3–20.

Friedman, M. and Schwartz, A.J. 1963. *A Monetary History of the United States, 1867–1960*. Princeton: Princeton University Press.

Gravelle, H. and Rees, R. 2004 [1981]. *Microeconomics*. Harlow: Pearson Education.

Humphrey, T.M. 1997. 'Fisher and Wicksell on the Quantity Theory'. *Federal Reserve Bank of Richmond Economic Quarterly* 83:4:71–90.

Newcomb, S. 1966. [1885]. *Principles of Political Economy*. New York: Harper's.

Robbins, L. 1998. *A History of Economic Thought: The LSE Lectures.* Oxford: Princeton University Press.

Schumpeter, J.A. 1956. [1951]. *Ten Great Economists: From Marx to Keynes*. London: Allen and Unwin.

Somerville, R.A. 2004. 'Changes in relative consumer prices and the substitution bias of the Laspeyres price index: Ireland, 1985–2001'. *Economic and Social Review* 35:1:55–82.

Tobin, J. 1987. 'Fisher, Irving' in *The New Palgrave: A Dictionary of Economics* 2:369–76.

Trautwein, H-M. 2007. 'Book review: Dimand & Geanakoplos (eds) 'Celebrating Irving Fisher'' *History of Political Economy* 14:3:607–11.

WAR GAMES: MILITARY STRATEGY AND ECONOMIC GAME THEORY

PETER DEVINE

Junior Sophister - Visiting Student

Confining the use of economic models to explaining economic phenomena ignores the insights such models can provide in a variety of other areas. In this paper, Peter Devine illustrates one such alternative application by examining how economic models based on game theory and guided by so-called 'Just War Principles' could help guide military policy intended to reduce civilian casualties in a counterinsurgency. He uses the current example of the war in Iraq as the basis of this study.

Introduction

In the current war in Iraq, many innocent civilians are being killed by coalition forces in the process of defending the Iraqi government and people from a guerrilla insurgency. While the goal of the counterinsurgency is to protect civilian lives, implementing policies to achieve that objective can be complex. Many factors affect how competing incentives interact and some outcomes can be counterproductive, causing even greater harm to civilians than would otherwise be the case.

Economic modelling based on game theory and guided by Just War Principles could be employed to more accurately characterize the interaction of competing incentives and, in the process, help counterinsurgent commanders to make informed decisions that greatly decrease civilian casualties. While a model that completely characterizes the dynamics of the Iraq War would be extremely complex, a simplified model that analyses the key trade-offs is still broadly useful. Furthermore, a simple model can be incrementally expanded and improved upon by including institutional effects that incorporate individual nuances of the situation.

Civilian Casualties are Unacceptable Losses

Since the 2003 US led invasion 655,000 more people have died in Iraq than would have otherwise (Burnham et al., 2006). Of those excess deaths, 31% were attributed to coalition forces. It can't be disputed that too many innocent civilians have been killed in the conflict for no obvious reason. As recently as October of 2007, Iraqi Prime Minister Nuri al-Maliki complained of the 'excessive force' used in a raid by American troops which killed 13 civilians, two of which were toddlers, and injured 69 others (Raheem and Kami, 2007). In fact, between January and May of 2007, US and NATO forces have killed more civilians in Afghanistan than the insurgent forces (USA Today, 2007). In reaction, Afghan President Hamid Karzai declared that US and NATO forces viewed civilian lives as 'cheap' (ibid). In many circumstances civilian casualties are extremely difficult, perhaps impossible, to avoid during military operations. However, soldiers and policy makers should aim to prevent civilian deaths. Innocent lives deserve that sincere effort.

The Competing Incentives of an Insurgent War

The problems associated with countering an insurgency fought within an urban environment are complex. Coalition forces currently engaged in Iraq are struggling with the insurgents' ability to fight while dispersed among the civilian population where a significant number of deaths of innocent civilians can be attributed to the counterinsurgency (O'Hanlon and Campbell, 2007; Burnham et al., 2006). Military strategy affects the duration of the conflict, the number of civilian casualties, influencing public opinion in favour of, or opposed to, the counterinsurgency (Patraeus and Mattis, 2006; Walzer, 2006). Decisions not only need to be effective in terms of overcoming the insurgency, they also need to protect the rights of civilians by abiding to the limits of waging a just war (Walzer, 2006).

The goal of the counterinsurgent force is to re-establish the legitimacy of the government by gaining the support of the local population. Minimization of civilian casualties is an essential aspect of achieving that goal and the conflict cannot be won if the local population feels an excessive number of innocent civilians are dying unnecessarily: 'The cornerstone of any [counterinsurgent] effort is security for the civilian population' (Patraeus and Mattis, 2006).

In every insurgent conflict there are three categories of civilians: a minority which actively supports the insurgents, a minority which actively supports the counterinsurgents and a neutral majority (ibid). It is the third group,

the majority, whose support must be won in order for the government to attain legitimacy. The only way to win the support of the majority is by providing for their security and establishing the rule of law (ibid). From this perspective, every counterinsurgent operation is won or lost by the counterinsurgent force's ability to protect the public. A successful counterinsurgency requires that combatant commanders and troops have the safety of civilians as their primary concern. To do so, the counterinsurgent forces have to assume greater personal and company risk in order to protect the civilian population in which they operate (ibid).

While avoiding civilian casualties is the primary focus, military commanders are still faced with difficult strategy decisions on how to accomplish that goal. Stringently avoiding any action that might harm civilians could paradoxically have the opposite effect and actually increase civilian casualties. For instance, a policy that prioritizes civilians' safety makes civilians valuable shields to hide behind. A policy that is not affected by the presence of civilians eliminates their value as shields but creates a moral dilemma. Even if the policy decreases the net civilian casualty rate over the course of the conflict, innocent civilians may still be harmed. This example demonstrates that in order to minimize civilian casualties it may actually be necessary to accept some level of endangerment to civilian lives. To do otherwise may cause harm to a greater number of civilians in the long run. However, any policy that causes increased civilian deaths in the short term in order to reduce civilian deaths over the long term risks causing the local population to lose sight of potential long-term benefits.

The Theory of Just War

It is warranted under the Theory of Just War to fight a counterinsurgency where civilian casualties occur so long as the act of war satisfies the criteria of the Double Effect Principle (Walzer, 2006). The Double Effect Principle outlines criteria under which an act that has unintended harmful effects is justified by the more significant and intended helpful effects. The Double Effect Principle declares that a good or helpful act which yields unintended, harmful effects (the double effect) is justified if the following four criteria are met:

1. The nature of the act is itself good
2. The intention is for the good effect and not the bad
3. The good effect sufficiently outweighs the bad effect to merit the risk of yielding the bad effect
4. The good effect is not a result of the bad effect.

Civilian casualties may be acceptable under the Theory of Just War so long as maximum effort is employed to protect against them. The Counterinsurgent field manual requires combat units to accept greater company risk in order to provide greater security for the public (Patraeus and Mattis, 2006).

How an Economic Model Can be Useful

Decision makers trying to choose between warfare strategies need tools for predicting outcomes arising from various courses of action. It is extremely difficult to determine the optimum level of force that will meet a military objective while protecting civilians to the maximum extent possible. The impact of a military action can be counterintuitive leading to bad decisions and lost civilian lives. An economic model that characterizes the competing incentives could be useful to evaluate courses of action against expected outcomes.

The complexities associated with implementing a Just War policy to counter an insurgency dispersed among a civilian population need to be better understood. Economic modelling and game theory can help study and characterize those complexities by supplying decision makers with tools to objectively evaluate various battlefield strategies against likely results (Osborne and Rubinstein, 1994; Gibbons, 1992; Straffin, 1993; Walzer, 2006; O'Brien, 1981; Yoder, 2001). The projection models proposed are based on existing economic/game theory and can be tailored to take into account the limits imposed by *'Jus in Bello'*.[1] For example, a model that characterizes the tendency of an insurgent force to fight dispersed within a civilian population can be structured as a two-player game in which the costs and benefits to each 'player' are modelled as a function of the level of force employed by the counterinsurgents and the dispersion level of the insurgents. Various military strategies could be analyzed in terms of predicted civilian casualties. The model's outcome projections represent one source of objective data points which decision makers could use to refute or support a proposed strategy.

[1] '*Jus in Bello'* is a term which describes fighting a war justly.

The Dispersed Insurgent Force Model

A Dispersed Insurgent Force Model could be developed to characterize the effects of increased aggression against an insurgent force that is dispersed within a civilian population. The description below is just a starting point for such a model and is, therefore, highly simplistic. It is structured as a two-player game where each player seeks to maximize their own benefit to cost ratio. The counterinsurgent force is characterized by a cost-benefit curve in which increased aggressiveness against the dispersed insurgent force eliminates more insurgents yielding a benefit but at a decreasing rate. However, by increasing power the dominant counterinsurgent force also incurs a cost (civilian casualties) at an exponentially increasing rate. The net cost-benefit curve then shows a level of aggression at which an incremental increase in power would incur a cost greater than the benefit. Total cost (counterinsurgent) [1] is differentiated with respect to p to calculate marginal cost [2], where p is power and λ is dispersal:

$$p\lambda \quad [1]$$

$$\lambda \quad [2]$$

Total benefit [3] is differentiated to calculate marginal benefit [4], where α is the reciprocal of the duration:

$$(0.5p^{0.5}\alpha)/\ \lambda \quad [3]$$

$$\alpha/(4\ \lambda p^{0.5}) \quad [4]$$

By equating the marginal cost and marginal benefit we find the level of p at which benefit is maximized as discussed above:

$$\lambda = \alpha/(4\ \lambda p^{0.5}) \quad [5]$$

Solving for p:

$$p = \alpha^2/(16\lambda^4) \quad [6]$$

This yields the dominant counterinsurgent force's best response function, that is the level of p that will maximize the counterinsurgent's benefit for a given dispersal λ.

Similarly with the insurgent force, there is a cost-benefit curve for effectiveness and dispersal for a particular level of aggressiveness by the counterinsurgent force. As the dispersal rate increases, the insurgent forces gain protection. However, there is an inflection point where too great a dispersal reduces overall effectiveness because of a lack of organization and communication amongst the insurgents. Mathematically, total benefit [7] is differentiated to yield marginal benefit [8]:

$$\frac{1}{p} - (\frac{3p^2}{2} - \lambda)^2 \qquad [7]$$

$$3p^2 - 2\lambda \qquad [8]$$

By setting the marginal benefit equal to zero we find the level of dispersal λ which yields the maximum benefit to the insurgent force. Solving for λ:

$$\lambda = \frac{3p^2}{2} \qquad [9]$$

This yields the insurgent force's best response function; the level of dispersal λ that maximizes the insurgent's benefit for a given power setting p. Solving the two response functions yields an optimal Nash equilibrium. Substituting [9] into [6]:

$$p = \frac{1}{16(3p^2/2)^4} \qquad [10]$$

Note: We are assuming that duration α is not a factor – that is, the duration is not so long that the local public opinion turns against the counterinsurgency – α is set to 1.

$$p = 0.61369\ldots \qquad [11]$$

Substitute this value for *p* into [9] and solve for λ:

$$\lambda = 0.56492\ldots \qquad [12]$$

Therefore total civilian casualty rate in this case would be:

$$p\lambda = 0.34668\ldots \quad [13]$$

A field commander could use a model like this to predict the insurgents' dispersal rate given his ordered rate of aggression and thus calculate a projected civilian casualty rate.

Further Research Is Needed

In this example, only the most basic trade-offs are analyzed. In reality, fighting a guerrilla war is extremely complex and a model would have to include many variables. Extensive data would have to be collected and evaluated to construct accurate models for each player's costs and benefits. However, it is precisely because of its complex nature that such a model is needed to help manage the problem. Formal models help one comprehend how competing incentives interact. Although the problem being examined is complex, a simple model can be used to capture what are thought to be the key trade-offs. This framework could be potentially expanded to characterize institutional effects and nuances missed by the simpler model.

Furthermore the model needs to be extended from a one-shot simultaneous game to a sequential game. A counterinsurgent campaign is drawn out and highly adaptable by nature. There may be an element of simultaneous choice in each time period. However, a model needs to incorporate past reputation and follow sequential choice in order to be effective over the entire war.

An exciting possibility is that the amount of intelligence gathered on the insurgents' strategy and tactics could be a way to change the insurgents' cost-benefit curve. If so, the counterinsurgent force could significantly reduce civilian casualties by making choices that decrease the insurgents' dispersal within the civilian population.

Civilian support for counterinsurgent operations could be a free rider problem. While the civilian population may determine that it would be better off if the insurgent force were eliminated, no individual may be willing to take the first action against the insurgents because of fear of retribution. Increased security would result in higher net social benefit, but some will be harmed more than others in the process of opposing the insurgents. To incorporate this in the model, further research is required to analyse how the local government and the counterinsurgent forces would deal with this market failure.

Finally, how the duration of the conflict factors into the model needs to be better understood. In the model presented, duration is factored out for simplicity.

In reality, duration could be a significant factor. If the insurgency lingers on for an extended period, the indigenous population could begin to resent the counterinsurgents and their support for the counterinsurgency could erode.

Bibliography

Brown, D. 2006. 'Study Claims Iraq's 'Excess' Death Toll Has Reached 655,000'. *The Washington Post*. 11/10/2006:A12.

Burnham, G., Lafta, R., Doocy, S., and Roberts, L. 2006. 'Mortality After The 2003 Invasion Of Iraq: A Cross-sectional Cluster Sample Survey'. *The Lancet* 368:1421–28.

Gibbons, R. 1992. *A Primer in Game Theory*. New Jersey: Princeton University Press.

USA Today. 2007.' Afghan civilians reportedly killed more by U.S., NATO than insurgents'. 24/6/2007.

O'Brien, W. 1981. *The Conduct of Just and Limited Wars*. New York: Praeger Publishers.

O'Hanlon, M. and Campbell, J. 2007. 'Iraq Index: Tracking Variables of Reconstruction & Security in Post-Saddam Iraq'. *The Brookings Institution*. Viewed at: http://www.brookings.edu/iraqindex

Osborne, M. and Rubinstein, A. 1994. *A Course in Game Theory*. Cambridge: MIT Press.

Patraeus, D. and Mattis, J. 2006. 'U.S. Field Manual/Fleet Marine Force Manual'. FM 3-24/MCWP 3-33.5 Counterinsurgency. Washington.

Raheem, S. and Kami, A. 2007. 'U.S. military says kills 49 in Baghdad raid'. Reuters.com. 10/21/2007. Viewed at: http://www.reuters.com/article/newsOne/idUSCOL53898220071021

Straffin, P. 1993. *Game Theory and Strategy*. MAA.

Walzer, M. 2006. *Just and Unjust Wars* 4th ed. New York: Basic Books.

Yoder, J. 2001. *When War is Unjust* 2nd ed. Eugene: Wipf & Stock Publishers.

HOW VIABLE IS THE EUROPEAN UNION'S COMMITMENT TO BIOFUELS?

DAVID QUINN

Senior Sophister

Are biofuels the solution to our environmental woes? This certainly appears to be the policy being currently backed by the EU. A possible alternative to fossil fuels, biofuels have been touted as a means of reducing greenhouse emissions and providing increased energy security. David Quinn provides an objective assessment of EU biofuels policy. He notes the importance of subsidization in fuelling the current production drive, and the crucial role played by oil prices in determining relative competitiveness. He ultimately rejects biofuel policy as a sustainable long-term solution to Europe's energy requirements.

Introduction

> *'There has never been a better moment to push the case for biofuels. Crude oil prices remain high. We face stringent targets under the Kyoto Protocol, and the recent controversy over imports of Russian gas has underlined the importance of increasing Europe's energy self-sufficiency. Raw materials for biofuel production also provide a potential new outlet for Europe's farmers, who have been freed by CAP reform to become true entrepreneurs'*.[1]

We have moved past the point of legitimate dispute; climate change has finally been acknowledged as one of the greatest economic, social and environmental challenges of the 21st century. In recognizing the scope of this challenge, the nations of the world have begun to seek ways to reduce their negative environmental impact, and in particular to cut CO_2 emissions. Many activists feel

[1] Marianne Fischer Boel, European Commissioner for Agricultural and Rural Development, 8th February 2006 (European Commission, 2006)

that no economic power is doing enough to reduce these emissions and furthermore, that in the absence of a global carbon trading system, the effects on carbon reduction will be limited. One mechanism for decreasing CO_2 emissions that has been embraced by both developing and developed countries alike, is an expansion in the use of biofuels. Since 2000, global biofuel production has tripled from 4.8 billion gallons to over 16 billion gallons. Production is currently highly concentrated, with the U.S., Canada, Brazil and the EU contributing over 90% of the global supply of biofuels (OECD, 2006). In the face of unprecedented increases in oil prices, biofuels have been hailed as a key renewable substitute for oil. However, biofuels still make up only 3% of the supply of global transportation fuel, and production requires large amounts of arable cropland. These issues have raised growing concerns about the impact biofuels are having on commodity prices and the environment through increased demand pressures on agricultural land. With food prices on the rise, concerns about the ongoing viability of biofuels have entered into the public domain.

The purpose of this paper is to examine the economic viability of biofuels from a European perspective and discuss the current EU biofuels policy. The conclusion that emerges is that while biofuels do have the potential to help reduce CO_2 emissions, at current technology levels EU production is not viable. This paper begins by defining what exactly constitutes 'biofuels' and examines recent developments regarding biofuels in Europe. The next section will consider some of the recent criticisms raised against biofuels and the merits behind such criticisms will be explored. Finally, the future viability of biofuels is assessed, concentrating principally on the importance of oil prices and the potential for technological advances. The paper concludes with an outlook for the future.

What are Biofuels?

Biofuels can be defined as 'transportation fuels derived from biological (e.g. agricultural) sources' (IEA, 2004: 27). Biofuels can be produced from a variety of feedstocks, many of which are used in the agricultural food chain. They can be created in various liquid and gaseous forms. Ethanol and biodiesel are both liquid forms of biofuel. Ethanol is made from starchy and sugar crops, whereas biodiesel is made from vegetable oils derived from oilseed crops. Biofuels can also be produced from non-food organic materials including wood, cellulose, and waste materials. However, it is bioethanol and biodiesel that form the core of renewable transport fuels around the world (OECD 2006). The main advantage of these fuels is that they can be used either as pure fuels or blended with gasoline and diesel. In each case, up to 5% biofuel can be added to the mix

without any modifications to existing vehicle engines Brazil has been the world leader in bioethanol production for over 25 years. The EU currently leads the world in regard to biodiesel, which represents just under 80% of total EU biofuel production (OECD, 2006).

EU Policy

The European Union has set clear and ambitious targets for limiting CO_2 emissions. The EU has committed to a 20% reduction in greenhouse gas emissions by 2020 compared to the 1990 level, with an objective to reach a 30% international reduction through further negotiations (European Commission, 2007). Central to the success of this plan is the need to achieve a sustainable EU strategy on energy, particularly in the transport sector. The EU has set out a clear role for biofuels in achieving this goal. Biofuels contribute to EU policy objectives that include environmentally friendly security of supply, climate change commitments and the promotion of renewable energy sources (Commission of the European Communities, 2006b). Biofuels are also seen as a way to offer investment and employment opportunities for EU industries. In the agricultural sector, biofuels are expected to help smooth the sustained transition away from CAP payments. This last point is explicitly evident in CAP heavy countries, such as France, who have set even higher national biofuels targets.

The EU target for biofuels is to achieve a 5.75% share of the market for petrol and diesel in transport by 2010, with an interim target of 2% by 2005. However, these targets are not mandatory, with the EU instead opting for a review clause whereby the Commission was required to report on progress by 2006. This report estimated that in 2005 biofuels market share reached an estimate 1%, a doubling in two years (Commission of the European Communities, 2006a). While this can be deemed an impressive rate of progress, it is less than the 2% reference target, which was only reached by Germany (3.8%) and Sweden (2.2%). The report also acknowledged that it is unlikely the target for 2010 will be achieved, but indicated its desire to set a minimum target of 10% by 2020. In undertaking to reach these targets, biofuels development has been largely driven by financial incentives including subsidies, tax reductions and exemptions (Bamière et al., 2007).

While the EU should be praised for their proactive stance on CO_2 emissions, its current policy of expansion in biofuel production has raised a variety of concerns about the effect this expansion is having on agricultural land, food prices, and natural habitats. Questions have also been raised about the real reductions in CO_2 emissions that biofuels actually provide. These concerns were

acknowledged in January 2008 by the EU Environment Commissioner Stavros Dimas who commented that the EU has 'seen that the environmental problems caused by biofuels and also the social problems are bigger than we thought they were' (BBC News, 2007). The EU has indicated that it is prepared to abandon the targets if they turn out to be harmful to the environment or the disadvantaged in society. A number of the concerns that have been raised against biofuels in the recent literature are reviewed below.

Impacts of Biofuels on Climate Change

Biofuels have been presented as a key mechanism in the fight to reduce greenhouse gas (GHG) emissions within the EU. However, a number of recent studies have found that, when factoring in emissions from every stage in the supply chain, the energy balances of EU biofuels are less positive than evaluations originally suggested. These studies have estimated that the real reductions in GHG emissions may only be in the 25-30% range (Bamière et al., 2007; Farrell et al., 2006). Given that biofuels currently make up less than 1% of the EU's total energy consumption, the total effect on EU GHG emissions will be small. In addition, the current EU policies focus directly on production incentives and national supply targets, not on reducing GHG emissions. As a result, there has been no incentive to invest in the lowest GHG biofuel systems, only to raise production (Royal Society, 2008). Further concerns have also been raised about other negative environmental effects, including the effects of pesticides and nitrate pollution on water supplies, the loss of biodiversity, and increased deforestation, particularly in Brazil and Indonesia (Bamière et al., 2007).

Impacts of Biofuels on World Prices and the Poor

World feed grain prices are approaching all time highs (Kraemer & Schlegel, 2007). The International Monetary Fund calculated that world food prices increased by over 10% in 2006. This trend is expected to continue in the short and medium term. Food price increases can be attributed to factors including rising incomes, resource scarcity, lack of investment in agricultural productivity, and energy price increases. Numerous papers have documented the effect that biofuel production has had on food prices, particularly corn, wheat and soybean prices (Coyle, 2007; Bamière et al., 2007). The price of a bushel of corn has jumped from $1.86 at the end of 2005 to over $5 in 2008. Similarly, wheat is trading at over $10 a bushel, a doubling of its price in only 12 months. Futures contracts indicate that the high prices of both commodities will be maintained into the

immediate future.[2]

Some of the most stringent criticism of biofuels has been directed against their influence on food prices, and the knock-on effect this is likely to have on the worlds poor (Reuters, 2007; Washington Post, 2006). This argument was dramatically expressed by the UN Special Rapportuer, Jean Ziegler, who called the replacement of arable cropland with biofuels 'a crime against humanity' (BBC News, 2007). However, while rising food prices are a matter of major concern, there is another aspect of the debate that must be considered. Rising food prices may in fact offer an opportunity for low income countries to revitalise their agricultural and rural development. The agricultural policy of the developed world has had significant depressing effects on world food prices. Increases in food prices caused by biofuel production may just provide the impetus necessary to yield higher incomes and more jobs for food producers. 80% of food-insecure people live in rural areas. Given that there are powerful income and employment multiplier effects associated with agricultural-led growth, this effect could be substantial. High food prices have significant potential to reduce poverty and hunger across the world. However, policy is needed to ensure that poor families are provided with the opportunity to benefit from these high prices. This can be achieved by improving infrastructure and market coordination, encouraging contract farming and outgrower schemes, enforcing resource and land rights, and promoting competition in the marketing chain to ensure that higher prices really do benefit the poor (Matthews, 2008).

Impacts of Biofuels on Energy Security

The EU currently imports 50% of its energy needs. Without substantial domestic intervention, this is expected to rise to 70% in the next twenty to thirty years (Commission of the European Communities, 2006a). Given the uncertain nature both of oil supplies from the Middle East and gas from Russia, biofuels have been seen as key mechanism for ensuring EU energy security. However, EU analysis indicates that even if the EU biofuels policy is fully implemented, it will only provide a 3% decrease in fossil fuel imports (Commission of the European Communities, 2006b). Although this progress should be welcomed, it will not allow the EU to gain considerable self-sufficiently in terms of its energy needs. Furthermore, it is evidently unlikely that biofuel demand will be met from domestic supplies alone. While biofuel imports would allow the EU to diversify energy sources, they would not contribute to the goal of self-sufficiency.

---

[2] Figures taken from Chicago Board of Trade, 18/02/2008
Corn: http://www.cbot.com/cbot/pub/page/0,3181,1213,00.html
Wheat: http://www.cbot.com/cbot/pub/page/0,3181,1322,00.html

Is the Continued Level of EU Support Viable?

Given the criticisms outlined above, how viable is the EU's continued level of support for biofuels in the long run? To date, biofuel policy has been driven largely by political will, by both policymakers and the support the general public. This has meant that biofuel development has benefited from a large degree of subsidization. It is highly questionable whether this support will be sustained, especially considering the aforementioned concerns. Even the limited production of biofuels has had a dramatic impact on food markets, and the subsidization costs to Members States budgets have become significant. Certain Member States are making the transition away from tax exemptions to mandatory incorporation targets, and thus costs will eventually be passed to final consumers. As biofuel production expands towards the 5.75% target and then the 10% target, the costs of support will increase. The key challenge for biofuel producers will be whether or not they can justify continued support, either through valuation of their actual positive externalities or more practically, by competing with fossil fuels in terms of cost. The future of biofuels rests largely on two issues; the price of oil and the enhanced potential of second-generation biofuels.

The Economics of Biofuel Production

Oil prices are the single most important factor affecting the competitiveness of biofuels. Oil prices have tripled in value since 2003. We are entering a period of sustained high oil prices. In fact, oil futures prices indicate that oil is expected to remain on or above $70 a barrel.[3] This has had a huge impact on alternative energy sources, greatly enhancing their relative competitiveness. However, higher oil prices also lead to higher biofuel production costs; both in terms of higher feedstock prices and higher energy costs. Considering feedstocks make up a significant proportion of biofuel production costs, 80% in the case of EU biodiesel from rapeseed, this effect can be significant. The interconnectedness of oil prices, biofuel production and feedstock costs makes analyzing the competitiveness of biofuel production a formidable task.

Production costs for biofuels vary considerably across feedstocks and countries. However, the OECD (2007) estimated that Brazil has the most competitive biofuel production and can produce economically viable ethanol at around $39 a barrel. US Maize ethanol is competitive at $45 a barrel. EU production fairs less favorably, with wheat ethanol and sugar beet ethanol both competitive at just $100 a barrel. Biodiesel from rape oil fares marginally better

[3] NYMEX Crude Oil Futures, 18/02/2008. Viewed at:
http://futures.tradingcharts.com/marketquotes/index.php3?market=CL

at around $90 a barrel. At current oils prices, EU biofuel production is not competitive. This has important implications for EU biofuels policy. At current technology levels, biofuel production is simply not financially competitive.

Second-generation biofuels

Second-generation biofuels play a key role in EU biofuels policy. These are produced with ligno-cellulosic biomass, which utilises the entire plant. These second generation technologies avail of a greater amount of biomass resources from agriculture, forestry and waste materials, and are expected to achieve greater fuel production and lower GHG emissions. The most encouraging second-generation biofuel technology to date has been ligno-cellulosic processing from forest materials. Pilot schemes are currently in operation in Denmark, Spain and Sweden. Germany and Sweden are also currently piloting Fischer-Tropsch biodiesel and bioDME technology in demonstration plants (DG Agriculture, 2006)

However, these technologies are not yet commercially viable. The EU has indicated a firm commitment to research and development through its 1.75 billion 7th Framework research programme which gives high priority to research into next-generation fuels from biomass. However, there is still a great deal of technical and economic uncertainty. To date, studies assessing the impact of investment in second-generation biofuels have been inconclusive and we have no precise forecasts regarding second-generation biofuels (Bamière et al., 2007). One such study by the European Energy Agency (2006) concluded that the biomass potential is sufficient to reach the 2010 target and could increase to 16% of EU energy needs in 2030, even with strict environmental constraints. However, this investigation does not analyze the measures needed to ensure these constraints, and fails to provide an assessment of future technology requirements. Furthermore, the study makes a number of questionable assumptions about increases in productivity and agricultural liberalisation.

Conclusion

Despite the era of high oil prices, the future of biofuels remains uncertain. Current EU biofuel production is not financially or environmentally sustainable and there are lingering doubts over the economic viability of second-generation biofuels. However, biofuels currently remain the most obvious alternative to fossil fuels. As such, there is still a role for biofuel policy in the European Union. EU policy needs to shift focus away from supply-based targets towards policy that promotes the production of biofuels that are the most cost competitive and

environmentally sustainable. A reassessment of the strategic importance of first generation biofuels is necessary, and greater priority should be given to research into second-generation biofuels (Doornbosch & Steenblik, 2007). Further research is also needed to better estimate the real GHG and environmental effects of biofuel production in order to estimate the appropriate level of subsidy payments for environmental externalities.There remains a substantial need to improve current understanding of the effects that higher oil prices are going to have on biofuel production.

Developing countries may well have a key role to play in biofuel production. Greater research is required to assess the relative importance that biofuel exports might play in producing more efficient and less polluting energy sources, without damaging the ecological systems of producing nations. Efforts are required in order to lower trade barriers, and thus allow developing countries to take advantage of their comparative advantages in biofuel production (Doornbosch & Steenblik, 2007).

Finally, it is important to note that despite their promise, biofuels are only one of a wide array of policy options available for tackling climate change. Without serious global cooperation and commitment, the reduction in CO_2 emissions will be highly limited. Greater leadership is necessary in order to create market conditions that account for the real social cost of carbon. Without this, all efforts to tackle climate change will fall short. A carbon tax on energy sources that emit CO_2 pollution is perhaps the most effective option, as this would create a market system that adequately accounts for the cost of pollution. A possible consequence of this would be the encouragement and development of the most climate friendly biofuels, as well as making sectoral targets such as the percentage targets for biofuels irrelevant. As such, the EU's continued domineering focus on biofuels targets is misguided and efforts should be made to refocus on the most effective mechanisms of reducing CO_2, i.e. pricing carbon. The danger remains that the continued debate over biofuels will detract from this goal.

Bibliography

Bamière, L., Bureau, J.C., Guindé, L., Guyomard, H., Jacquet, F. and Tregeur, D., 2007. 'Prospects for Biofuels in the EU: Imports or Local Production?' *TradeAG* Draft Working Paper 2007.

BBC News. 2007. Biofuels 'crime against humanity'. 27/10/2007. Viewed at: *http://news.bbc.co.uk/2/hi/americas/7065061.stm*

Commission of the European Communities. 2006a. *Green Paper: A European*

Strategy for Sustainable, Competitive and Secure Energy. COM(2006)105 final,

Commission of the European Communities. 2006b. *An EU Strategy for biofuels*. COM(2006)34 final.

Coyle, W. 2007. *The Future of Biofuels: A global perspective*. US Department of Agriculture.

DG Agriculture: European Commission. 2007. *Biofuels in the European Union: An Agricultural Perspective*. Fact Sheet: Brussels.

Farrell, A. E. et al. 2006. *Ethanol can Contribute to Environmental and Energy Goals*. Science 311:506

IEA. 2004. *Biofuels for Transport – An International Perspective*. International Energy Agency, Paris.

Kraemer, A. and Schlegel, S. 2007. *European Union Policy on Bioenergy, Policy Brief*. German Marshall Fund of the United States, Washington, D.C.

Matthews, A. 2008. *Are Rising World Food Prices a Threat to the MDG Hunger Target?* MDG Lecture Series, Trinity College Dublin. 13/02/2008.

OECD. 2006. *Agricultural Market Impacts of Future Growth in the Production of Biofuels*. AGR/CA/APM(2005)24. Paris: OECD.

OECD. 2007. *Implications for agricultural markets and policies*. AES Conference on Bioenergy: London. 31/01/2007.

Reuters. 2007. *As biofuels boom, will more go hungry?* 07/03/2007. Viewed at: http://www.reuters.com/article/inDepthNews/idUSL0639150720070307

Royal Society. 2008. *Sustainable Biofuels: Prospects and Challenges*. London: Royal Society.

Washington Post. 2006. Storming the people to feed the cars. 10/09/2006. Viewed at: http://www.washingtonpost.com/wpdyn/content/article/2006/09/08/AR2006090801596.htlm

IS THE EUROPEAN UNION'S SINGLE FARM PAYMENT TRULY DECOUPLED?

ANN STILLMAN

Senior Sophister

Any form of subsidization has a distortionary effect on market outcomes. Ann Stillman evaluates the most recent EU agricultural reforms, assessing the status of the Single Farm Payment (SFP) and its relationship to production decisions. She acknowledges that the direct connection has been severered. However, by examining the indirect distortionary effects, she concludes that the Single Farm Payment cannot yet be regarded as fully decoupled from production.

Introduction

The issue of decoupling agricultural support from production has in recent years gained much momentum in both policy and academic circles. In the European Union (EU), this has coincided with the design and implementation of a series of Common Agricultural Policy (CAP) reforms which have effectively marked a shift from previously dominant market price-support centred policies towards a more decoupled form of support. The 2003 Mid-Term Review (MTR) of the Agenda 2000 Reforms, proposed by then Commissioner for Agriculture and Rural Development Franz Fischler, is the latest and perhaps most radical of these. Specifically, while the 1992 MacSharry and subsequent Agenda 2000 reforms initiated a process of partial decoupling—whereby coupled market-price support was progressively reduced while compensatory direct payments to farmers were increased—the MTR replaced all premia and arable aid payments with a consolidated Single Farm Payment (SFP). The SFP is tied, not to production volumes or yields, but to average historic payments received by farms in the base period 2000-2002, contingent only upon retention of entitlement acreage and a number of 'cross-compliance' criteria.

This paper will seek to assess the extent to which the SFP is truly decoupled. The question is underpinned by substantial ambiguity regarding the appropriate definition of 'decoupled' support. Therefore, the analysis will open

with a discussion of the relevant definitional issues. It will then go on to evaluate the SFP against this backdrop of conflicting conceptualisations, in order to determine how effectively the post-Luxemburg Agreement incarnation of the CAP meets its stated objective of decoupling agricultural support from production decisions.[1]

Definitional Ambiguities: varying perspectives on the true nature of decoupled support

The debate on policy decoupling is characterised by a multiplicity of differing interpretations of what constitutes a 'decoupled' payment. Any critical assessment of the extent of decoupling must therefore explicitly consider these varying perspectives.

A first notable point in which definitions differ is in their adoption of *ex ante* versus *ex post* approaches. The former, often associated with policy-makers, focuses on the eligibility criteria attached to payments: policies are deemed to be decoupled if eligibility depends on a fixed historical base period, and direct payments are financed by taxpayers, not linked to current prices, production or factor use. This *ex ante* perspective has been articulated, for instance, by Burfisher and Hopkins [2003]. It contrasts with the latter approach, typically favoured by agricultural economists, which targets not policy design, but policy effect, and hence considers policies to be decoupled from production only if they have no influence on the relative prices and quantities of agricultural outputs produced or inputs used to produce them.

Considering the notion of decoupling from an *ex post* economic perspective, Cahill (1997), drawing on a large body of earlier writing (e.g. Andersson, [2004]; Baffes, [2004]; Beard and Swinbank, [2001]; Breen et al., [2005]; Swinbank et al., [2004]), has made an influential distinction between what he terms full decoupling and the less restrictive concept of effective full decoupling—respectively placing emphasis on adjustment or equilibrium. Fully decoupled policies, in his view, are those that '[do] not influence production decisions of farmers receiving payments, and that permit free market determination of prices' (Cahill, 1997: 351). Effectively fully decoupled policies, on the other hand, are those that result in a level of production no greater than the equilibrium level of output that would be observed in the absence of the policy for any/all types of crop. This dual view of decoupling has been frequently cited

[1] We will for the most part limit the scope of our discussion to only those reforms established in the original 2003 agreement.

in the literature surrounding the subject (e.g. Andersson, [2004]; Baffes, [2004]; OECD, [2001]; Rude, [2007]).[2] These definitions are not, however, without their limitations. In particular, the notion of fully decoupled policy measures is useful as a theoretical construct, but given the difficulty of establishing specific supply and demand schedules empirically, is of little practical utility for policy analysis. Effective full decoupling overcomes this issue to some extent—however, its asymmetrical focus solely on policy effects resulting in excess output, rather than both positive and negative production effects, is problematic.

Furthermore, any attempt to operationalise these definitions must establish some basis for quantifying decoupling. Cahill thus introduced the Degree of Decoupling index, defined as one less the ratio of the production effects of a policy over the production effects of a positive price change of identical magnitude (Cahill, 1997; OECD, 2001). This, in turn, raises a number of questions. In particular, the index measures the extent of decoupling against the benchmark of a fully coupled policy, which is taken to be market-price support. This is somewhat paradoxical—policy measures are evaluated according to their relatively inferior production-distorting effects rather than compared to unregulated equilibrium outcomes.[3]

There is also the question of whether policy measures should be evaluated in isolation or as a complete policy package. To this effect, the OECD (2001: 12), following Cahill (1997), has concluded that 'the policy package matters', and therefore investigations of decoupling should consider the specific effects of single measures as well as the overall effects of bundles of measures. Gohin et al. (2000) have further substantiated this analytically, concluding that policy impact varies depending on whether measures are implemented jointly. Finally, the discussion so far has centred on the production effects of agricultural policy; however, any analysis must not neglect potential policy impact on consumption, as both of these aspects can directly contribute to trade-distorting effects (OECD, 2001).

Elements of *ex post* economic conceptions have been combined with *ex ante* policy design focused elements into a more pragmatic legal or

[2] In graphical terms, a policy that is fully decoupled according to Cahill's terminology would have no effect on the shape of supply and demand functions, while one that is effectively fully decoupled would have no influence upon market equilibrium outcomes, although it might affect the shape of the supply and/or demand schedules and hence result in different responses to exogenous shocks (OECD, 2001).

[3] This can be seen as reflecting the path dependency phenomenon proposed by Kay [2003] and discussed by Swinbank et al. (2003). In this vein, Pierson has argued that 'public policies [are] not only outputs of but important inputs into the political process, often dramatically reshaping social, economic, and political conditions' (Pierson, 1993: 595).

administrative definition that is appropriate for prescriptive policy purposes. Under the World Trade Organisation (WTO) guidelines, agricultural policy measures have been categorised according to the extent to which they are 'decoupled' from production. Policies that comply with so-called 'Green Box' criteria are exempt from support reduction commitments under WTO rules provided they 'have no, or at most minimal, trade-distorting effects or effects on production' (URAA Annex 2, Art. 1). In addition to this fundamental *ex post* requirement, a number of *ex ante* conditions must be met in order for measures to qualify for Green Box status, among which are requirements for clearly defined eligibility criteria with a fixed historic base period, payments that are unrelated to the type or volume of agricultural commodity produced, market prices or inputs used in any other than the base year, and that payments are not contingent upon any production whatsoever (URAA Annex 2, Art. 6). These definitional issues are a persistent undercurrent to both policy-making and policy analysis, as will be apparent throughout the following section.

The Single Farm Payment: testing theoretical hypotheses on production effects

Ex Ante Limitations of the SFP

The CAP reforms of the last two decades have been hailed as indicative of a fundamental paradigm shift in European agricultural policy-making (Daugbjerg, 2003). The historic state-assisted paradigm, founded on the principle that: 'First, the agricultural sector contributes to national policy goals and therefore merits special attention; and, second, the price mechanism is a suboptimal means of achieving an efficient and productive agricultural sector' has, according to this view, given way to a market liberal paradigm in which market forces are the prime determinants of agricultural supply (Coleman et al., 1997: 275). However, an *ex ante* approach to reviewing the SFP already reveals several contradictions in the Commission's stated objective of enabling 'complete farming flexibility increasing market orientation' (Commission, 2002: 19). Most obviously, concessions made to member states in negotiations mean that the final legislation passed deviates significantly from the original proposal in permitting partially coupled support to be retained for some commodities and livestock (Binfield et al., 2004). In addition, land employed for the cultivation of fruit and vegetables was initially excluded from entitlement to the SFP (Commission, 2003). This incomplete decoupling is likely to affect farmers' resource allocation decisions, shifting production from fully decoupled towards partially decoupled commodities. Moreover, the persistence of import tariffs, export subsidies, and

intervention price guarantees undermines the community's commitment to market liberalisation (Matthews and Dixon, 2006).

In addition to these flagrant inconsistencies, labelling the SFP as 'decoupled' ignores the fundamental fact that, while no production is required, payments remain linked to the primary factor of production, land. Historical area-based payments, therefore, can impact on land purchase and rental costs, as the benefits of entitlement to direct payments is capitalised into land values. This, in turn, may distort production through increased barriers to entry and effects on farm profitability (OECD, 2005a). A substantial body of literature has emerged to support the significant long-run effects on land values of 'decoupled' direct payments (Bhaskar and Beghin, 2007). Similarly, coupling to land reduces incentives for farmers to exit the agricultural sector—mounting a barrier to exit—as this would entail the loss of SFP income. According to Gohin et al. (2001, cited in Rude, 2007: 7), 'If the amount of the direct payment exceeds the loss associated with a particular productive activity, then there may be a cross subsidization effect that will keep that producer in business'. This cross subsidization effect is consistent with research by Chau and De Gorter (2005). Although it applies primarily where payments are conditional on production, Rude (2007) has argued that cross-compliance requirements may have a comparable effect. Coupling the SFP to land thereby curtails the structural adjustment which is supposedly at the centre of agricultural policy reform objectives.

Furthermore, the so-called cross-compliance criteria—which make receipt of the SFP contingent on fulfilment of a number of statutory standards including maintenance of land in agricultural condition, environmental conservation, and public, plant and animal health and welfare requirements (Commission, 2003)—may influence production in so far as they restrict the scope of farmer decision-making in various ways (OECD, 2005b). Therefore, rather than reflecting what Isabelle Garzon has branded as a shift from a dependent to a multifunctional paradigm, this could be interpreted as an attempt to disguise and preserve the prevailing system of dependence and state assistance by restricting movement towards truly decoupled support measures (Potter and Burney, 2002).

To overcome these residual coupling mechanisms, early proponents of decoupling advocated the establishment of a transferable bond scheme. Perhaps most notably associated with Tangermann (1991), this would provide an annuity to farmers to compensate them for the reduction in coupled market-price support and would be entirely unconditional—decoupled from land and free from compliance requirements. Because bonds could be traded in a secondary market, the value of bonds would not filter into land prices or restrict farmers' decisions to hold land, produce, or otherwise adjust to market conditions. The Fischler

reforms took a small step in this direction by permitting farmers to sell their payment entitlements. However, because once sold entitlements had to be again linked to an equivalent area of eligible acres, payments remained coupled to land (Tangermann, 2003).

Lastly, even in its legal/administrative sense, characterisation of the SFP as decoupled is questionable. The MTR proposal stated that the SFP would 'provide a major advantage within the WTO, since the Green Box compatibility of the scheme will help secure these payments in an international context' (Commission, 2002: 19). However, the Upland Cotton Dispute brought before the WTO Dispute Settlement Body against the US has set an important precedent. In this case, the Appellate Body ruled that direct payments and other benefits to cotton farmers could not be classified as permissible decoupled payments under Annex 2 of the Green Box (Oxfam, 2004). While reforms to the fruit and vegetable regime in 2007 should in theory avert similar cases being brought against the EU's SFP Scheme, Swinbank (2007) has questioned whether cross-compliance criteria infringe upon the scheme's eligibility for Green Box status. Swinbank and Tranter (2005) have similarly concluded that the SFP may not fit within the Green Box due to the conditionality placed on eligibility. Therefore, departing from an *ex ante* perspective already reveals a number of limitations of the SFP as a truly decoupled measure.

Risk Effects

Wealth and Insurance Effects

Furthermore, if these limitations of the MTR are ignored and it is assumed for the sake of argument that the SFP is indeed entirely independent from market prices, with no conditionality attached to payment eligibility, several indirect production effects may persist. Foremost among these is the potential for 'decoupled' lump sum payments to distort producer decision-making via wealth or insurance effects. Assuming farmers are rational utility-maximising agents that display constant relative risk aversion (CARA), or decreasing absolute risk aversion (DARA), then lump sum direct payments such as the SFP will result in a fall in the coefficient of absolute risk aversion, increasing farmers' appetite for risk. Similarly, the SFP will smooth income variability in relative terms resulting in an insurance effect which, under the same conditions, will reduce absolute risk aversion (Bhaskar and Beghin, 2007). These effects have perhaps most notably been modelled by Hennessy (1998), who concludes that uncertainty regarding future agricultural outcomes e.g. yields, market prices, given DARA preferences, depresses production. Wealth effects of support policies will therefore have a

positive impact on production. Hennessy (1998: 49) also demonstrates that, 'As with the wealth effect, the insurance [effect] fortifies the producer with the confidence to increase production when increased production is associated with increased risk.' Further research is consistent with these findings (Sckokai and Moro, 2006). Indirectly, wealth effects may also influence farmers' labour allocation decisions, both in terms of labour/leisure and on-farm/off-farm labour choices (Bhaskar and Beghin, 2007).

Dynamic Effects

Investment Effects

Wealth and insurance effects may furthermore be significant in a dynamic setting. Specifically, SFP may result in higher current savings and hence greater scope to increase future investment. To this effect, the OECD (2005b) has estimated that area payments, e.g. the SFP, have significant effects on investment, with insurance effects having a greater impact than wealth effects. Moreover, where credit constraints act as a limiting factor on farmer decisions, the capitalisation of direct payments into land values previously discussed, as well as the guaranteed income from direct payments will improve the creditworthiness of farmers and facilitate access to credit thus enabling increased investment (Goodwin and Mishra, 2006). These investments may increase production not only directly, if they are in yield-enhancing technologies, but also indirectly, if by lowering unit production costs they increase production incentives. In so far as greater asset holdings and values increase the overall wealth of the farm sector, this positive feedback perpetuates wealth effects over time (Westcott and Young, 2002).

Expectations

Another key dynamic consideration involves the role of farmer expectations regarding future policy-making on production decisions. The underlying rationale is as follows: even if direct payments are decoupled from current production and based on a fixed historical period, farmers may anticipate, for instance, a future updating of the base period or re-coupling of payments so that decisions to reduce production or planted area now would result in a loss of entitlements to future support payments. Kydland and Prescott (1977: 486) have suggested that discretionary policy-making yields sub-optimal social outcomes due to their time inconsistency because 'current decisions of economic agents depend upon expected future policy, and these expectations are not invariant to the plans selected.' This is consistent with the findings of empirical and experimental evidence cited by Bhaskar and Beghin (2007). In addition, survey

data on Irish farmers presented by Hennessy (2004) and Breen et al. (2005) indicates a reluctance to alter production patterns in the farm sector in response to the MTR, which may well stem from expectations regarding future policy changes. Thus, farmers' reservations regarding the longevity of current policies or expectations of future policy shifts may hamper adjustment decisions and thus negate the effects of policy reforms.

Conclusion

The above discussion has demonstrated that the SFP, in both policy design and effect, falls somewhat short of being truly decoupled Limiting our assessment just to an *ex ante* conceptualisation of the term, it is already apparent that the SFP is by no means the 'no strings attached' payment it has been presented as. While the direct link of support to production has been removed, the payment remains coupled to land as well as to a number of conditions that place restrictions on farmers' freedom to respond to market incentives. Moreover, assuming the scheme did achieve its objectives of decoupling from an *ex ante* perspective, substantial evidence suggests that even unconditional lump-sum payments will have an indirect distortionary influence on production, via wealth, insurance, investment, and expectations effects. Thus, as Spriggs and Sigurdson have put it, 'the only truly decoupled program that there is,' is 'a program to eliminate subsidies completely' (1988, as cited in Baffes, 2004: 4).

Bibliography

Baffes, J. 2004. 'Experience with Decoupling Agricultural Support'. *Rod Ziemer lecture*. University of Georgia. 14/04/2004.

Bhaskar, A. and Beghin, J. 2007. 'How coupled are decoupled farm payments? A review of coupling mechanisms and the evidence'. *Working Paper #07021*. Iowa State University.

Binfield, J., Donnellan, T. Hanrahan, K., Hart, C. and Westhoff, P. 2004. 'CAP Reform and the WTO: Potential Impacts on EU Agriculture'. *Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting*. Denver: Colorado. July 1-4, 2004.

Breen, J., Hennessy, T. and Thorne, F. 2005. 'The effect of decoupling on the decision to produce: An Irish case study.' *Food Policy* 30: 2: 129-144.

Cahill, S. 1997. 'Calculating the rate of decoupling for crops under CAP/oilseeds reform,' *Journal of Agricultural Economics* 48: 349-378.

Chau, N. and de Gorter, H. 2005. 'Disentangling the Consequences of Direct Payment Schemes in Agriculture on Fixed Costs, Exit Decisions, and Output'. *American Journal of Agricultural Economics* 87: 5: 1174-1181.

Coleman,W., Skogstad, G. and Atkinson, M. 1997. 'Paradigm shifts and policy networks: cumulative change in agriculture'. *Journal of Public Policy* 16: 3: 273–301.

Commission of the European Communities. 2002. Brussels, 10/7/2002 [COM(2002) 394 final].

Commission of the European Communities. 2003. Brussels, 21/1/2003. [COM(2003) 23 final].

Daugbjerg, C. 2003. 'Policy feedback and paradigm shift in EU agricultural policy: the effects of the MacSharry reform on future reform'. *Journal of European Public Policy,* 10: 3: 421 – 437

Gohin, A, Guyomard, H. and Le Mouël, C. 2000. 'Measuring the Degree of Decoupling of Alternative Internal Support Policy Instruments: The Green Box Issue'. INRA-ESR, Rennes. *Paper presented at the IX European Congress of Agricultural Economists*. Warsaw, August 24-28, 1999.

Goodwin, B. and Mishra, A. 2006. 'Are 'Decoupled' Farm Program Payments Really Decoupled? An Empirical Evaluation'. *American Journal of Agricultural Economics* 88: 1: 73–89.

Hennessy, D. 1998. 'The Production Effects of Agricultural Income Support Policies under Uncertainty' .*American Journal of Agricultural Economics* 80: 46–57.

Hennessy, D. 2004. 'Farm Level Adjustment in Ireland Following Decoupling'. *Paper prepared for the 78th annual Agricultural Economics Society Conference*.

Kydland, F. and Prescott, E. 1977. 'Rules Rather than Discretion: The Inconsistency of Optimal Plans'. *The Journal of Political Economy*, 85: 3: 473-492.

Matthews, A. and Dixon, N. 2006. 'Impact of the 2003 Mid-Term Review of the Common Agricultural Policy'. *Quarterly Economic Commentary* Ireland: ESRI.

Organisation for Economic Co-operation and Development. 2001. 'Decoupling: A Conceptual Overview.' Paris.

Organisation for Economic Co-operation and Development. 2005a. 'Decoupling: Illustrating Some Open Questions on the Production Impact of Different Policy Instruments.' Paris.

Organisation for Economic Co-operation and Development. 2005b. 'Modeling the Impact of Agricultural Policies on Farm Investments Under Uncertainty: The Case of the CAP Arable Crop Regime.' Paris.

Oxfam. 2004. 'Dumping: the Beginning of the End? Implications of the Ruling in the Brazil/US Cotton Dispute'. Oxfam Briefing Paper 64.

Pierson, P. 1993. 'When effect becomes cause: policy feedback and political change'. *World Politics* 45:4:595:628.

Potter, C. and Burney, J. 2002. 'Agricultural multifunctionality in the WTO: legitimate non-trade concern or disguised protectionism?' *Journal of Rural Studies* 18:35–47.

Rude, J. 2007. 'Production Effects of the European Union's Single Farm Payment'. *Canadian Agricultural Trade Policy Research Network* Working Paper 6.

Sckokai, P. and Moro, D. 2006. 'Modeling the Reforms of the Common Agricultural Policy for Arable Crops under Uncertainty'. *American Journal of Agricultural Economics* 88: 1: 43–56.

Swinbank, A. and Tranter, R. 2005. 'Decoupling EU Farm Support: Does the New Single Payment Scheme Fit within the Green Box?' *The Estey Centre Journal of International Law and Trade Policy* 6: 1: 47-61.

Swinbank, A. 2007. 'Decoupling EU Farm Support: WTO Challenges to Mr. Fischler's CAP.' *Based on a paper presented to the conference Agricultural Policy Changes: Canada, EU and the World Trade Organisation*. University of Victoria: Victoria, BC, Canada. 13-15 September 2007.

Tangermann, S. 1991. 'A bond scheme for supporting farm incomes' in Marsh, J. and Green, B. 1991. *The Changing Role of the Common Agricultural Policy: The Future of Farming in Europe*. London: Belhaven.

Tangermann, S. 2003. 'Decoupling and the Way Forward for the CAP' in OECD. 2003. *CAP in Hand: Reflections on a Decade of CAP Reform*. Brussels.

Westcott, P. C., and Young, C. E. 2002. 'Influences of Decoupled Farm Programs on Agricultural Production'. *Presented at Free Trade Area of the Americas, The WTO, and New Farm Legislation: Responding to Opportunities and Challenges*. San Antonio, Texas. May 23-24, 2002.

Uruguay Round Agreement on Agriculture (URRA), Annex 2, WTO.

SPLENDID ISOLATION: BRITAIN'S PAST, PRESENT AND FUTURE IN RELATION TO THE EUROPEAN MONETARY UNION

DAVID MADDEN

Junior Sophister

Was it politics or economics that kept Britain from joining the single currency? David Madden presents a brief monetary history of Europe since the First World War, and examines the question of Britain's dedication to the European project. Although it may have made economic sense to remain outside in the past, he suggests that the time is now right for Britain to join the single currency. However, he concludes that the eurosceptical attitude of the British public remains too great an obstacle, leaving economics to take a backseat to politics.

Introduction

Britain's committment to the ideals of the European project has always been in question. This lack of enthusiasm for integration was evident in the 1991 decision to obtain an opt-out clause from the Maastricht Treaty, which stated that it was not bound to join the monetary union. This allowed it to keep full control over monetary policy and avoid having to adopt the euro as its currency. Through insisting upon this opt-out clause, Britain has undoubtedly distanced itself from the rest of Europe. Of course this is nothing new in British history. In the late 19th century the British government, under the premiership of Benjamin Disraeli, followed a foreign policy known as 'splendid isolation' in relation to Europe. This involved keeping Britain firmly out of European affairs by not committing her to any of the alliances of great powers that were being formed at the time. It is this essay's intention to examine why 'splendid isolation' has become the defacto policy of the British government in relation to the European Monetary Union. The history of monetary integration in Europe will be discussed, followed by and examination of the rationale behind the formation of the EMU. This essay seeks to discover why Britain opted out, and questions whether Britain was correct in this decision by analysing developments since 1999. Finally there

is a brief discussion of whether we can expect to see Britain's 'splendid isolation' from the EMU continue into the future.

A Brief History of Monetary Integration in Europe

The idea of monetary integration between European countries only became plausible following the collapse of the Bretton Woods system in 1973. Under this arrangement, gold was the ultimate source of value, with the dollar being the only currency directly tied to it. All other currencies were defined in terms of the dollar. High levels of inflation during the late 1960's and early 1970's forced the USA to suspend the dollar's gold convertibility in 1971 and devalue it by 10%. This led to a wave of exchange rate realignments designed to restore order to the system. Unstable currency markets and a lack of credibility in the system led to the decision in 1973 to abandon the 'fixed but adjustable' principle. This effectively meant that each country was free to choose its own exchange rate regime.

Europe's early and rapid reaction to this new system charted the path that would lead to a monetary union three decades later (Baldwin and Wyplosz, 2004, 2006). The 'European Snake' was formed to curb intra-European exchange rate fluctuations. It quickly collapsed due to the failure of a number of countries, including Britain, to keep inflation in check. What followed was the establishment of the European Monetary System (EMS) in 1979. At the centre of the EMS was the Exchange Rate Mechanism (ERM) whose stated purpose, according to the Council of Ministers, was to establish 'a greater measure of monetary stability in the community' through the joint management of fixed and adjustable exchange rates.

During the first ten years of the EMS, Eurpean inflation rates diverged markedly and realignments were a frequent occurrence. The transparent process of exchange rate realignments meant that the foreign exchange markets in EMS countries were subject to frequent speculative attacks. The only way to prevent such turmoil and uncertainty was to reduce inflation differentials between countries. Thus more and more countries took to emulating the monetary policy of Germany, as the largest country with the lowest rate of inflation in the EMS. In effect, the Bundesbank became the centre of the European monetary system. This however didn't prevent another series of speculative attacks in 1992 and 1993 which nearly destroyed the EMS and forced the margins of fluctuations to widen to ±15%. The EMS is still in existence but it now acts as a half-way house for member states seeking to gain entry into the EMU.

The EMS had always been seen as a preliminary stage in the EC's

progress towards monetary unification (Davies, 1989). The fall of the Berlin Wall in 1989 had given the then President of the European Commission, Jacques Delors, the perfect opportunity to propose forming a monetary union. After lengthy negotiations the Maastricht Treaty was agreed upon in 1991, which committed the EU to forming a monetary union by 1999 and adopting a single currency by 2002. The establishment of the European Monetary Union represented a massive pooling of sovereignty between the eleven original member countries. It required members to transfer their control of national monetary policy and interest rates to the European Central Bank in Frankfurt, as well abandon their national currencies in favour of a new single European currency. There existed both an economic and a political rationale for the formation of the EMU. Economic motives included:

> 'enhanced monetary stability, improved spatial and industrial allocation of resources, boosted competition because of transparent prices, deepened integration, reduced transactions costs, access to wider markets, gains from economies of scale and from trade, as well as gains that come from harmonisation of policies' (Jovanovic, 2005: 119).

Political impetus centred on France's deep-seated fears that the reunification of Germany would lead to a return to German militarism. In their opinion this could only be prevented through an increase in the forces that tied Germany to Europe.

Britain's Opposition to Monetary Integration

Britain, along with other EU15 members Sweden and Denmark, has thus far chosen to opt out of EMU. But this is nothing new, it's attitude towards monetary integration in Europe has always been lukewarm at best. Although a member of the EMS, it refused to participate in the ERM until 1990. The main reason for this lay in the fact that Britain's rate of inflation was far above that of the European average. Had Britain entered into the 'fixed but adjustible' system of bilateral parities, sterling would have been forced into numerous exchange rate realignments in order to maintain trade competitiveness within the system.

However, by 1985 the rate of inflation in Britain was temporarily below the European average and it seemed to many, including the Chancellor of the Exchequer, that the time was right for sterling membership (Davies, 1989). Yet there were two obstacles to such a course of action. Firstly, sterling's position as a petro-currency remained important, prompting the government to argue that it

would be both awkward and inappropriate to join the system at a time of greatly fluctuating oil prices (Davies, 1989). Secondly, most calculations showed that sterling was slightly overvalued against the Deutschmark. It was thus thought foolish to lock in such a high real exchange rate at a time when North Sea oil production was reaching it's peak, and when a downward trend appeared to be developing in the real price of energy (Davies, 1989). Both of these obstacles, coupled with Mrs. Thatcher's innate fear of being seen to cede British sovereignty to Brussels, succeeded in keeping Britain out of the ERM until 1990. By this stage it had become a political necessity for Britain to join, as failure to do so would seriously damage its chances of playing a constructive and active part in the negotiations surrounding discussions on monetary union (George,1990, 1994, 1998).

It was generally considered by observers that John Major had achieved quite a diplomatic and political success by obtaining the British opt-out from monetary union. Public opinion in Britain was firmly behind such a decision due to commonly held anxieties relating to both economic performance and sovereignty. Historically Britain had tended to be out of line with it's European neighbours when it came to economic cycles. There existed a fear in Britain that it might find itself in recession while the rest of it's partners were overheating, and consequently would be subject to a restrictive and wholly inappropriate monetary policy (Davidson, 1996). Counter-arguments pointed out that the impending formation of the Single Market would smooth out such cyclical disparities. However for most Britons forming a monetary union before the Single Market had even proved to be successful was too big a risk to take.

Abandoning the power to introduce a competitive currency devaluation was another source of disquietude in Britain. If the rate of productivity growth in Britain were to be consistently lower than in the other EMU nations, or if increases in unit wage costs were to be consistently higher, it would be unable to compensate by devaluation and its loss of competitiveness would be reflected in a politically intolerable increase in unemployment (Davidson, 1996). Other economic concerns included the politically independent nature of the proposed European Central Bank. The opinion of many was that such independence could only work well in a country such as Germany where there was a deep political and social commitment to the objective of low inflation (Davidson, 1996). In contrast, the population in Britain were less committed to low inflation and placed far greater emphasis on political accountability. The notion of having an unelected bureaucracy controlling monetary policy did not sit well on the British conscience. However, such economic concerns were minor in comparison to the widely held fear that joining the EMU would be the first step towards becoming part of a political union in Europe. The British public fiercely opposed any

perceived attempts to erode it's proudly held sovereignty. In their eyes, adopting the euro in favour of the pound and transferring control of monetary policy to Frankfurt, would be the beginning of the end for British independence. Such narrow and uncompromising views were fuelled to a large degree by the virulent eurosceptical media.

Was Britain Correct to Opt Out of the EMU?

The answer to this question seems to be yes, but there are a number of mitigating factors to take into account. Since 1999 economic growth in Britain has far exceeded that of the Eurozone as a whole. Average growth in Britain has been 2.8% compared to 2.14% for the Eurozone (European Commisssion 2004). Between 2000 and 2005 interest rates in Britain averaged 4.87% compared to 4.48% for the Eurozone (ibid). Britain's inflation levels were much lower in the same period averaging 1.35% in comparison to 2.17% for the Eurozone as a whole (ibid). These simple statistics highlight a number of significant issues. Firstly Britain's economic cycle does indeed appear to be out of line with the rest of the Eurozone. Since the establishment of the EMU, it's growth has far outstripped the rest of Europe. In the event that it had transferred control of monetary policy to the ECB, it's economy would have been far more likely to overheat due to the unsuitably low interest rates that have existed in the Eurozone over th relevant period. However, it is worth noting that the British economy is one of the largest economies in the EU. Had it been part of the Eurozone, it would have had considerable influence over decision making in the ECB. This would have allowed it to push interest rates up above the level they were at between 2000 and 2005. Furthermore, the experience of Eurozone countries, such as Ireland, which experienced above average growth demonstrates that the negative effects of an inappropriate monetary policy are not nearly so great as first thought.

Has Britain's Attitude Towards the EMU Softened?

Although Britain does not apeear to have suffered economically by remaining outside the single currency, there exists a 'wait-and-see attitude' towards the EMU. When he was appointed Chancellor of the Exchequer in 1997, Gordon Brown devised five economic tests for assessing whether it was in Britain's interest to join the Eurozone. These tests related to the following: economic convergence with the Eurozone, sufficient flexibility to adapt, impact on jobs, effect on financial services and influence on foreign direct investment. After

publishing 18 detailed studies and a lengthy assessment on British membership of the Eurozone in 2003, the Treasury's answer was 'not yet' (Jovanovic, 2005). It was found that the convergence and flexibility criteria were not fulfilled, that the investment and financial services tests were met, and the fifth test would be met when the first two were met.

Many commentators feel that the question of Britain remaining outside of the EMU ultimately hinges on the flexibility test. Christopher Swann wrote: 'Flexibility would be Britain's insurance policy inside the Eurozone' (The Financial Times, June 2003). Considering Britain has one of the most flexible labour markets in the EU, should it not be sufficiently flexible to join? Christopher Swann summed up the argument against this assertion by writing: 'If the Eurozone suffered an economic shock and other nations adjusted more slowly than the UK, the Treasury argues, monetary policy would become inappropriate for the UK' (ibid). Mr. Brown clearly believes that Britain will only be able to join the Eurozone when all of the EMU member countries are able to able to adapt effectively to shocks in the EU.

A second source of hesitation, apparent since 1999, involves the exchange rate issue (De Grauwe, 2007). The pound sterling has experienced strong cyclical movements against the euro. Between 1995 and 2000 it experienced a strong appreciation relative to the euro. From 2000 to 2006 this was partially reversed. Joining the EMU when sterling was in such a strong position relative to the euro, would have saddled Britain with low competitiveness and would have put downward pressure on economic growth.

Is the Time Now Right for Britain to Join?

After assessing Britain's past and present in relation to the EMU, it is now necessary to discuss the possibility that Britain will join at some point in the near future. It can be argued that the time is right for Britain to take its place as one of the dominant economies within the EMU. The success of the Single Market in facilitating labour and capital flows within the EU has meant that Britain is more economically tied in than ever to rest of Europe. This should lead it's asynchronous economic cycle to closer conformity to those of the rest of Europe. Labour market reforms in both France and Germany should also enhance the flexibility of the Eurozone as a whole in respond to shocks. Furthermore, the existence in Britain since 1997 of an independent Central Bank should facilitate the transfer of monetary sovereignty to the ECB, and lessen the culture shock to the British electorate. Sterling's depreciation in recent months in relation to the euro is another reason why it is in Britain's interest to join. Since November 2007

the pound has fallen by nearly 9% against the euro- a rate of decline not far off that experienced during sterling's enforced exit from the ERM in 1992, when it dropped 11% against the Deutschmark. By joining now, Britain could enhance its competitiveness and boost economic growth.

If all of these reasons were not sufficient to suggest that the time is right for Britain to join, the global credit crunch of recent months provides further justification for this argument. Among central banks, only the Bank of England took no contingency measures in August when the crisis erupted. This seriously calls into question its ability to adequately deal with the effects of global shocks on the British economy. Its failure to provide sufficient liquidity to the money markets has seriously undermined confidence in the stability of the banking system. It was this lack of liquidity that caused Northern Rock to get into so much financial trouble. In contrast, the European Central Bank provided weekly liquidity injections into money markets. This undoubtedly helped a number of European banks weather the storm more favourably.

Conclusion

There would undoubtedly be costs to Britain joining the EMU at this moment in time, but such costs would far outweigh the benefits. However, it is highly unlikely Britain will join the EMU in the short term. Anti-European sentiment still remains strong amongst Britons, clearly illustrated by the public controversy surrounding the ratification of the Lisbon treaty. Despite all of this, splendid isolation cannot last forever. The need for Europe to stand up and speak with one voice in world affairs means that EU integration will continue apace. Britain's splendid isolation from the EMU will have to cease if it wants to be at the centre of an increasingly influential and economically powerful EU.

Bibliography

Baldwin, R. and Wyplosz, C. 2006. *The Economics of European Integration*. London: McGraw-Hill Education.

Davidson, I. 1996. *European Monetary Union: The Kingsdown Enquiry: The Plain Man's Guide and the Implications for Britain*. London: Macmillan.

Davies, G. 1989. *Britain and the European Monetary Question*. London: Institute for Public Policy Research.

European Commission, 2004. 'European Economy: EMU after 5 years. Special Report'. No.1.

de Grauwe, P. 2007. *Economics of Monetary Union*. Oxford: Oxford University Press.

George, S. 1998. *An Awkward Partner.* New York: Oxford University Press.

Jovanovic, M. 2005. *The Economics of European Integration: Limits and Prospects*. Cheltenham: Edward Elgar.

The Financial Times. 2003. *Euro Decision the Five Tests: Brown reluctant to accept similar risks for labour market as eurozone states*. 10/06/2003.

UNDERSTANDING LIQUIDITY CRISES: THE THEORY OF HYMAN MINSKY

MICHAEL BRACKEN

Senior Sophister

In this topical paper Michael Bracken endeavours to clear up the confusion surrounding the term 'liquidity', a concept that is frequently misinterpreted, misused and misunderstood. Using the theory of Hyman Minsky, he outlines the crucial role played by liquidity in the stability of the global financial system. In this context, there is a brief discussion of the Long-Term Capital Management debacle and the origins of the current market turmoil. Like Keynes before him, Minsky emphasized the unstable nature of a system built on expectations of future cash flows, and advocated government intervention in order to stabilize the system in times of crisis. In terms of crisis prevention, the conclusion is somewhat negative: crisis is inevitable under a capitalist structure; the best we can hope for is to minimize the damage.

Introduction

Our financial system was struck by a crisis again this summer and liquidity was at its heart. This paper begins by defining the term 'liquidity' and goes on to examine why it plays such a fundamental role in the stability of the global financial industry. By using the theory of Hyman Minsky, this paper aims for an understanding of why financial crises keep reoccurring. The paper concludes by presenting Minsky's proposals on minimising the damage of such crises.

Defining Liquidity

> 'Central bankers have drowned the world in it, oil producers are awash with it, while an excess of it distorts everything from treasury yields to the copper forward curve. Yet overnight this all-powerful force can vanish, causing markets to tumble…

> Liquidity in its first, narrow, definition is an important economic concept. But in its more fashionable second usage, liquidity is too, well, wishy-washy, to be useful'(The Financial Times, August 2007).

George Orwell, in a 1946 essay[1], said that the inaccurate use of written words eventually confuses our thinking and this is evidently the case with liquidity, a term which has had its popular meaning expanded greatly by sloppy usage. The above excerpt from the FT gives an insight into the importance of liquidity in financial markets and the development of its 'fashionable second usage', which uses 'excess liquidity' as a catch-all phrase to denote: 'loose central bank policy rates, broad money supply growth, aggressive lending to private equity, yen borrowing and even the growth of debt derivative products' (ibid). The grouping of such diverse phenomena under one term is worse than useless; it clarifies nothing and only serves to confuse. It is therefore preferable to look beyond this 'fashionable' usage to the term's historical origins.

It was centuries ago that realising the assets and discharging the liabilities of a firm in distress was first referred to as liquidation. Then in the nineteenth century bankers began to refer to government bills as self-liquidating. However, an understanding of the notion of liquidity as applied to assets is a relatively recent concept, emerging from the work of J.M. Keynes: he first discussed it in his Treatise on Money (1930) and brought it to prominence through his central role in the writing of the Macmillan Report[2]. It was established as a principal contribution of Keynes with the publishing of *The General Theory of Employment, Interest and Money* in 1936 and the widespread discussion of his liquidity preference theory (Hicks, 1962).[3]

Keynes' first definition of liquidity in the Treatise has proven remarkably robust and is what the FT calls the 'first, narrow definition' of liquidity: 'Bills and call loans are more liquid than investments, i.e. more certainly realisable at short notice without loss' (Hicks, 1962: 4). A liquid asset will have high marketability and high capital certainty; the asset can be realised relatively promptly for close to its full market price (Moore, 1968). The most liquid asset then is clearly cash as it can be realised instantly for its entire value and other assets can be ranked in liquidity relative to it.

---

[1] *Politics and the English Language* (Orwell, 1946)

[2] More formally known as the British Treasury Committee on Finance and Industry Report (1931)

[3] Keynes' liquidity preference theory established a link between liquidity and risk by redefining interest as a reward for surrendering liquidity rather than a reward for saving.

It is important to have a clear understanding of liquidity as it is, generally speaking, central to all financial crises; people wish to sell assets but they cannot do so at a speed and price that is favourable to them. It is fundamentally what led to the insolvency of Northern Rock in August 2007, as it could find nobody to rollover its commercial paper in the money markets, and goes back to John Law's Mississippi System. Kindleberger (2005: 28) uses the wonderful German word Torschlusspanik[4] to describe the resulting panic selling that a liquidity decrease generates. But it must be noted that liquidity is therefore simply a descriptive term, to describe it as the cause of anything is to put the cart before the horse. For example, during the endgame of Law's system, to say that a decrease in liquidity led to a fall in prices is a tautology – to understand the liquidity crisis we must know the causes that have led to this lack of liquidity.

The Theory of Hyman Minsky

The most compelling analysis of the causes of such liquidity crises is to be found in the work of Hyman Minsky. Minsky was an economist who, while too mild to be called an iconoclast, was certainly an anachronism in the post war era. His aspiration to expose the instability inherent in the capitalist system was conspicuously in opposition to the spirit of the 1950's: America was experiencing a prolonged period of growth and low inflation, while Howard Macmillan was telling the British people 'You've never had it so good'(Bellofiore, 2001: 35). Minsky was also a follower of Keynes and a macroeconomist in an era when the Chicago school and rigorous mathematical microeconomic underpinnings to theories[5] were beginning to dominate the landscape. It is unsurprising then that Minsky's views on financial crises were largely ignored by his contemporaries and have only recently begun to receive widespread attention, with Kindleberger adopting them for his prominent work on the subject.

Minsky believed that 'the essential critical flaw in capitalism is instability' (Minsky, 1982: 86). He opposed the neoclassical belief that if undisturbed from outside[6] the economy would naturally find an equilibrium that is consistent with full employment. Minsky maintained that instability is endogenous to system, a view he arrived at from focusing on elements of Keynes ignored by the neoclassical synthesis, namely: the pricing of capital assets, the

[4] 'Door-shut-panic'. Investors crowd through the door before it shuts.

[5] As exemplified by Arrow and Hahn's (1971) general competitive equilibrium model.

[6] The textbook example of such an exogenous shock is a change in the supply of oil.

importance of capitalist financial institutions and the uncertainty surrounding decision-making. In short it was a more 'real world' view of the economy, and the opposite of the neoclassical approach which Minsky jokingly referred to as 'the economics of capitalism without capitalists, capital assets and financial markets' (Minsky, 1986: 120).

This Keynesian view of the importance of capital assets and companies as the buyers of these assets led Minsky to begin his analysis in a money economy with sophisticated financial institutions. Money here was the product of financial interrelations, a financing veil that hides the ultimate wealth owners from the financial assets.[7] This money was created by banks in the process of financing investment, and so central banks could not control the effective money supply. Any attempt to do so would simply lead to a new innovation. Henry Simons describes this process as: 'the reappearance of prohibited practices in new and unprohibited ways… it seems impossible to predict what forms the evasion might take' (Minsky, 1982: 71). Minsky, appropriately for a student of Schumpeter, recognised that 'the history of money is the history of innovation around legislation to use existing money more efficiently and create new substitutes' (Kindleberger, 2005: 58).

This view of money necessarily equates money with purchasing power, and so, in the words of Mill, that a man's purchasing power: 'consists, first, of the money in his possession; second, of the money at his banker's, and all the other money due to him and payable on demand; thirdly of whatever credit he happens to possess' (ibid: 61). Minsky focused on the pro-cyclical nature of the credit supply as the source of economic instability, reflecting his other great influence (along with Keynes): Henry Simons. Simons would have supported the Baron de Rothschild's view that: 'If speculators could find unlimited credit, one can't tell what crises would ensue' (ibid: 177) and so called for a 100% reserve backed currency. This proposal was rejected by Minsky due to its knock-on effects on investment. However, he successfully went on to unite these aspects of Simons with Keynes and create his own unique view of the economy with his Financial Instability Hypothesis.

[7] Owners of capital assets generally borrow to obtain them. Once this debt is repaid they owe money to the true owners of the assets who have a claim on this money. Money separates the true wealth owner from the financial asset.

The Financial Instability Hypothesis

> 'A capitalist financial system will be capable of generating the signals that induce an accelerating desire to invest and of financing that accelerating investment' (Minsky, 1982: 279).

Minsky begins his model with an economy that has had a cyclical past but is now growing steadily. Here two particular time-series are important: expectations of future cash receipts that have yet to be realised i.e. expected profits, and past cash payments that have given rise to current liability structures. Liabilities are innate to any capitalist system as 'entering and repaying debts are the essential processes of capitalism' (ibid: 72) due to the need for external financing of investment. Minsky splits liability structures in the economy into three distinct types based on the two time-series:

1. Hedge Financing: Where expected future cash flows are large enough to pay both the interest and principal on a firm's debt. The greater the equity in the liability structure, the greater the chance the unit is engaged in hedge financing.

2. Speculative Financing: Where expected future cash flows are only large enough to cover the interest payments on a firm's debt. The unit is left open to refinancing risk, as it must rollover the principal in the money markets.

3. Ponzi Financing: Expected future cash flows are not even large enough to fully cover interest payments on debt. This type of financing is even more precarious than speculative, as debt must be issued rapidly to cover interest payments.

The economy continues on a path of steady even growth for some time, but this stability bears the seeds of instability. Looked at from a Wall Street or City boardroom, the best performing units during this period are the indebted units, their leveraged positions allowing them to realise greater profits. The optimistic forecasts of the past have been validated and high liquidity preferences now look overcautious. Firms recognize the opportunity for greater future profits from investing today and so views on acceptable debt structure change.[8]

[8] Money is held because of uncertainty. When uncertainty declines, due to changing

In colloquial terms, greed begins to outweigh fear and expectations change accordingly. There exists little reason to be cautious in such an environment; management that cannot match the returns of levered competitors will be quickly ousted by shareholders. Indeed it can often be rational to follow the irrational mob; Kindleberger reports on a banker who purchased £500 of South Sea Company stock during the bubble, saying: 'When the rest of the world are mad, we must imitate them in some measure' (Kindleberger, 2005: 37). This mixture of greed and a herd mentality is neatly encapsulated in a quote from Citigroup CEO Chuck Prince in Summer 2007: 'When the music stops, in terms of liquidity, things will be complicated. But as long as the music is playing, you've got to get up and dance. We're still dancing' (The Guardian, 2007). Prince was forced to resign less than four months after making this comment upon revelations of Citi's exposure to potential write-downs of $11bn on illiquid securities.

Now a liability system begins to emerge that Simons calls: 'a mass of current obligations and a shoestring of equity' (Friedman, 1969: 85). Asset and stock prices boom due to the increase in investment and the availability of credit on favorable terms. Lenders use financial innovation to increase the credit they can generate from their capital, also being driven by greed as they see faster acting rivals increase profits. This boom may continue for some time, with the number of 'speculative' and 'Ponzi' units increasing as lenders offer better and better terms in the battle to win profitable market share. Kindleberger notes that: 'money always seems free in manias' (Kindleberger, 2005: 9) However, the entire edifice of debt is built on unstable foundations, conditional on expectations of future cash flows being realised.

An event Minsky calls a 'displacement' will inevitably occur altering these expectations as: 'the nature of the bubble is it that eventually it will be pricked and then as with a child's balloon the air may escape sharply' (ibid: 91). This event may be relatively minor, a bankruptcy perhaps or some similar event that leads to a small sell off, but the size is disproportionate to the change in expectations. Some borrowers' indebtedness now begins to look large relative to their potential income and they may decide to sell assets to meet the shortfall. Lenders will now be reluctant to rollover 'speculative' and 'Ponzi' debt and the credit supply dwindles. As asset prices fall, brokers will ask for greater margin payments, resulting in bankruptcy in some cases. All of this leads to further asset sales. The resulting fall in asset prices shunts more units into the 'speculative' and 'Ponzi' categories and the problem continues. Liquidity preferences increase, there is a change in acceptable liability structures, and unless action is taken, a

expectations, it makes sense to move some holdings of liquid assets into higher yielding capital assets.

resulting fall in investment and corporate profits. The economy may then enter a recession.

Empirical Views of Minsky

If, as Friedman claimed: 'theory is to be judged by the predictive power for the class of phenomena which it is intended to explain' (Friedman, 1953: 8) rather than the complexity of the mathematics of the assumptions, then Minsky's Financial Instability Hypothesis is a very good theory indeed. A definitive account of the Minsky theory applied to the history of financial crises can be found in Kindleberger (2005). In this paper, two particular instances of financial crises are discussed: Long-Term Capital Management and the current 'Credit Crunch'.

Long-Term Capital Management:
The crisis of Long-Term Capital Management (LTCM) effectively showcased the inadequacies of modern portfolio theory and highlighted the importance of Minsky's unique focus on liquidity. The facts surrounding the infamous hedge fund are widely known. The fund essentially made arbitrage bets on similar securities, shorting an over valued liquid bond and buying a similar undervalued illiquid one, on the assumption that spreads would narrow as the securities moved back to the equilibrium of the capital markets line. Due to the minute differences in value, LTCM used huge levels of debt to magnify returns.[9] In the view of modern portfolio theory the firm was relatively safe as it was engaged in a number of such trades and so had diversified away risk.

However, due to the Asian and Russian financial crises, the markets experienced a flight to the quality of US government bonds and LTCM's spreads widened instead of narrowing. The fund's brokers asked for more margin payments and LTCM found itself unable to borrow to meet these payments. However, the fund's positions were still profitable in the long run, evidenced by the fact that the banks that were forced to bail LTCM out later made a profit on the equity stake they received for their trouble. Modern portfolio theory sees long-run profitability as a guarantee of short-run finance, but in practice this is clearly not the case. Minsky's theory, of course, is based on practice and we clearly see that LTCM was exposed to a change in liquidity preferences, something which modern portfolio theory abstracts from.

[9] See Kindleberger (2005: 67)

The Current Credit Shortage

The continuing relevance of Minsky's theory can be demonstrated by analyzing the ongoing events from Summer 2007 in the financial markets, which closely mirror his model of the economy. A period of low base interest rates and a stable economic environment after the September 11th terrorist attacks encouraged more debt heavy liability structures to emerge. This could be seen in a number of ways in the financial markets. Private equity firms like Blackstone and Kohlberg Kravis Roberts went through a self proclaimed 'golden age', selling larger and larger amounts of debt to fund more audacious takeovers. Investor appetite for this type of debt appeared to have no limit. Banks began to offer so-called 'covenant-lite' loans, in an effort to stay competitive.[10] In America, home loans were offered to so-called 'ninjas', people with no income, jobs or assets. In many cases these loans were self-assessing. In short, credit terms were loosening, views on acceptable debt structures were changing and an edifice of debt was emerging.

The displacement occurred on the 9th August, when the market was simultaneously hit by two pieces of bad news. The ECB lent the sum of €95bn to lower inter-bank lending rates and the investment bank BNP Paribas halted withdrawals from three mortgage-backed security (MBS) funds as they could not give the assets a 'fair value'. It was clear that liquidity preferences had changed; MBS had become illiquid and banks were unwilling to lend cash to each other. The first bank run in the British Isles in more than a century occurred on Northern Rock, a bank whose business model epitomised the credit mania[11]. Minsky's model could almost be an ex post description of the crisis.

Conclusion

Although Minsky presented a clear exposition of the inevitably of liquidity crises in capitalist economies, did he provide us with a solution? He certainly advocated a role for the government in stabilising a crisis. Minsky was a sup-

[10] That is, loans with few protective covenants, which require the borrower to provide certain information, not take certain courses of action etc. Loans effectively became more like bonds.

[11] Instead of building a retail deposit base, like most banks, Northern Rock mainly funded itself through the cheaper wholesale markets, leaving it open to changes in market sentiment.

porter of the classic Bagehot 'lender of last resort' role for the central bank: to put a floor on asset prices and the government deficit spending to counter the fall in investment. In regards to crisis prevention, he failed to provide a practical solution. Whereas Simons believed that there was some stable form of capitalism possible through reform of the financial structure, and Keynes believed that government intervention could accomplish this (Friedman, 1969: 81-95), Minsky accepted neither. According to Minsky, instability is an intrinsic part of capitalism. He states that:

> 'The history of money, banking, and financial legislation can be interpreted as a search for a structure that would eliminate instability. Experience showed this search failed and theory indicates a search for a permanent solution is fruitless' (Minsky, 1986: 314).

However with proper institution and policy we can control this volatility, 'we can, so to speak, stabilise instability' (ibid: 9).

Bibliography

Bellofiore, R. and Ferri, P. 2001. *Financial Keynesianism and Market*

Instability: The Economic Legacy of Hyman Minsky Vol. 1. Cheltenham: Edward Elgar Pub.

The Financial Times. 2007.*Defining Liquidity*. 11/08/2007.

Friedman, M. 1969. *The Optimum Quantity of Money and Other Essays*. London: Macmillan.

Friedman, M. 1952. *Essays in Positive Economics*. .Chicago: University of Chicago Press.

The Guardian. 2007. *When the Music Stops*. 06/11/2007.

Hicks, John. 1962. 'Liquidity'. *The Economic Journal* 72:288:787-802.

Kindleberger, C. 2005. *Manias, Panics and Crashes: A History of Financial Crises*. Basingstoke: Palgrave Macmillan.

Minsky, H. 1982. *Inflation Recession and Economic Policy*.. Brighton: Wheatsheaf.

Minsky, H. 1986. *Stabilising an Unstable Economy*. London: Yale University Press.

Moore, B. 1968. *Introduction to the Theory of Finance*. Collier-Macmillan.

THE CONTRIUBTION OF BLACK, MERTON AND SCHOLES TO FINANCIAL ECONOMICS

IAIN GERARD NASH

Senior Sophister

The introduction of the Black-Scholes Formula revolutionised the world of financial economics and lead to the creation of new fields, new markets and new types of securities. In this paper, Iain Nash examines the impact which the formula has had on the realm of economics and in particular the options market. Through his analysis of the combined contribution of Fisher Black, Robert Merton and Myron Scholes to the creation of this formula he clearly highlights the fundamental role such pioneers and visionaries play in the progression of a discipline.

Introduction

> *'Robert C. Merton and Myron S. Scholes have, in collaboration with the late Fischer Black, developed a pioneering formula for the valuation of stock options. Their methodology has paved the way for economic valuations in many areas. It has also generated new types of financial instruments and facilitated more efficient risk management in society'.*[1]

The contribution of Black, Merton and Scholes to the study of financial economics is a broad and profound one. They are most commonly associated with the development of the first complete options pricing formula. However, the impact of this formula extended to the creation of: 'a new field within finance, known as derivatives and offered a new perspective on related areas including corporate finance, capital budgeting and financial markets and institutions' (Arrow 1999: 230). Arrow also discusses how the work of Black, Merton and Scholes has influenced the development of mathematical and computer science,

[1] 1997 Nobel Prize for Economics – Press Release
Available at http://nobelprize.org/nobelprizes/economics/laureates/1997/press.html

as well as revolutionising private industry practice (Arrow, 1999). Few economists can claim to have had such an impact upon such a diverse range of fields, in both practical and academic terms.

A Brief History of Options

The theory of option pricing can be traced back to Bachelier (1900), whose PhD thesis presented an asset pricing model described by Duffie (1998) as quite similar to that of the Black-Scholes approach. This was one of the first papers which dealt with what would now be described as 'derivatives'. Arrow describes the thesis, while having somewhat flawed economic and mathematical inclusions, as being one of the primary motivations behind the work of Black, Merton and Scholes (Arrow, 1999). There had been work conducted on the area by Sprenke [1961], Ayres [1963], Boness [1964], Samuelson [1965], Baumol, Malkiel and Quandt [1966] and Chen [1970]. However, none of the formulae developed by these authors were 'complete' as they all contained arbitrary and non-estimable parameters and as such, they had no practical value; however, they did help create insights which were used by Black and Scholes in their paper (Black and Scholes, 1973).

Elton et al. define an option as 'a contract entitling the holder to buy or sell a designated security at or within a certain period of time at a particular price' (Elton et al 2007: 576). Two of the most common and simple options are 'calls' and 'puts'. A call gives the holder the right to purchase a security at a predetermined price, while a put gives the holder the right to sell a security at a fixed price. Many complicated options can be considered (and valued) as combinations of various calls and puts.

Options can be further broken down into 'European' and 'American' options. These are not actually based on geographical factors but rather European options can only be exercised at a set, predetermined maturity date, while American options can be exercised at any time up until the maturity date. The Black-Scholes model is only able to calculate European style options (Black and Scholes, 1973). For the calculation of American options, other pricing methods such as the Binomial Options Pricing Model must be used.[2]

Options are traded in two forms of markets: organised exchanges such as the Chicago Board Options Exchange and the New York Board of Trade, or in

[2] A good treatment of this model is given in Cuthbertson & Nitzsche (2001).

Over the Counter (OTC) markets. Organised exchanges follow standard terms and are centralised, while OTC markets are composed of networks of institutions, traders and investment banks and there is no central locus. Options markets differ from other forms of markets in that they are 'zero supply' markets (Arrow, 1999). Each trade must have a seller and a counterparty, and for each option sold there must be an option purchased. OTC markets are more dynamic then organised exchanges, as the lack of standard trading practices allow the investment banks and other types of specialist institutions to create custom products for clients. In the OTC market, it is also possible to create a 'one sided' trade if the broker is able to construct a synthetic portfolio (which are discussed later in this paper) as a counterparty.

Black-Scholes Formula

The history of the development of the Black-Scholes formula is covered extensively in Black (1989) but a brief outline is given below. In essence, Black and Scholes combined the intuition of the Capital Asset Pricing Model (CAPM) with stochastic processes. This lead to the development of the Partial Differential Equation (PDE) that was to become the Black-Scholes formula. Black had established the initial form of this equation in 1969; however, he was unable to solve it at the time.

The solution to Black's PDE was the result of both mathematical and theoretical foresight. Black had teamed up with Myron Scholes in M.I.T. and they had begun working to solve the equation. However, it was Robert Merton who had developed a model using arbitrage for solving hedged portfolios. This model allowed Black and Scholes to solve the model differentially, and thus the famous model was born.

The intuition behind the model is based on the premise of creating a synthetic hedged portfolio to determine the price (premium) of put and call options. We follow the process as put forward by Arrow (1999: 233):

1. A call option is taken on an underlying stock:
   As the underlying stock price rises, the value of the call rises as it is more likely to lie above the strike price of the call.

2. A short position is then taken to offset the call:
   As the underlying price of the stock rises, the short position will decrease in value. This will partially offset the call. Naturally, this process works in reverse if the price of stock

falls, with the call option now offsetting the increased value of the short sale.

3. The positions are modified to create a perfect hedge:
The exact holding of the call and short strategies are modified to create a net return of zero. Any change in the price of underlying stock will be countered by either the call or the short sale.

This allows the calculation of the unique arbitrage free price of the portfolio and the change in the price of the option as a function of the change in the price of the underlying security. Creating a synthetic[3], perfectly hedged portfolio allows the PDE to be solved. What is special about the Black-Scholes formula is that it only depends on factors which are easily observable and available in the market place. Arrow lists these as: the strike price of the option, the current date, the current stock price, the risk free rate and the volatility of the underlying stock which is measured by its standard deviation (Arrow, 1999). Other models developed at the time depended upon factors that cannot be estimated such as the expected return of the asset or its risk premium.

The Black-Scholes Formula is given below, as presented by Elton et al. (2007: 592):

$$C = S_0 N(d_1) - \frac{E}{e^{rT}} N(d_2)$$

$$d_1 = \frac{\ln(S_0/E) + (r + \frac{\sigma^2}{2})t}{\sigma\sqrt{t}}$$

$$d_2 = \frac{\ln(S_0/E) + (r - \frac{\sigma^2}{2})t}{\sigma\sqrt{t}}$$

Where r is the risk free rate, C is the current value of the option, S_0 is the current price of the stock, E is the exercise price of the option, $e = 2.7183$, t is the time remaining before the expiration date as a fraction of the year, σ is the standard deviation of the stock, $ln(S_0/E)$ is the natural logarithm of S_0/E and $N(d)$ is the

[3] These synthetic portfolios are theoretical constructs rather then actual products. This is proved by contradiction, as if it were possible to easily create these portfolios, people would hold these rather then options.

value of the cumulative normal distribution.

Thanks to this formula, for the first time people were able to easily and accurately calculate options prices. Given that the formula was published around the same time that the Chicago Board Options Exchange (CBOE) opened, the Black-Scholes formula was a corner-stone in the development and transformation of the options market from a thinly traded, inactive and illiquid market to one which traded over 944 million options in 2007.[4]

The Impact of the Black-Scholes Model

Duffie notes how the Black-Scholes model is taught to every MBA student, most graduate and many undergraduate students who study either finance or economics (Duffie, 1998). Many people attribute the explosive growth of options markets to the model. It is generally acknowledged that the Black-Scholes formula was the catalyst which led to the development of 'derivatives', a whole new field in finance. While some people attribute the development to the options pricing papers which preceded the 1973 paper, the advent of the formula allowed the derivative market to develop as traders and brokers were now able to accurately price and measure risk (Arrow,1999; Konishi and Dattareya, 1996).

One of the most basic (and widely used) manifestations of the Black-Scholes model is the ability for speculators to 'bet' on the direction of equities. By using either puts or calls, speculators are able to profit by correctly estimating which direction a stock is going to move. By executing these transactions through options markets a speculator can realise the following benefits:

1. Lower cash outlay:
   By purchasing an option, a speculator does not have to actually own the underlying security. Given that option prices are often only fractions of the cost of the security, the speculator can make the same bet many times for the price of the underlying security.

2. Short selling:
   Until the advent of options markets, a speculator was unable to profit from the expected depreciation of a stock. However, with the advent of derivatives a speculator can

[4] http://www.cboe.com/AboutCBOE/History.aspx

now profit by either using a put or a futures contract.

3. Hedging:
Investors, firms and other market participants are able to protect themselves from being exposed to risk associated with the movement of equities by using derivatives.

Transactions such as these are now commonplace in the market but did not exist at the time when Black, Merton and Scholes were conducting their research. The Black-Scholes model also allowed traders to work out the markets perception of a stock volatility through the process of Implied Volatility. Implied Volatility is a dynamic measure and is: 'a matter of turning the option on its head and finding out what other traders and market-makers think volatility should be' (Sutton, 1990: 36). Essentially, Implied Volatility allows speculators to see how the market has viewed past volatility and determine how this view has changed over the sample. The results from this measure differ from the historical pricing of volatility, a more rudimentary and simplistic method of analysing the volatility of a stock. Kim et al. (2007) conclude that Implied Volatility provides an unbiased forecast of future volatilities and furthermore, that implied volatility is much more closely related to the volatility of excess returns then that of gross returns when the interest rates are highly volatile.

Other Contributions

The Black-Scholes formula also had a major impact on the field of financial risk management and in the valuation of corporate liabilities. Duffie (1998) discusses how the practice of financial risk analysis has cloned the method of the Black-Scholes formula:

> 'The idea that the option can be priced by finding a trading strategy that replicates its payoff is frequently used to hedge a given security, or even to hedge a given cash flow that is not traded as a security. If one is to receive an untraded option payoff, for example, the risk inherent in that payoff can be eliminated by selling the replicating strategy previously described' (Duffie, 1998: 419).

This practice has become particularly widespread in the investment banking community and allows them to create packages containing virtually any combination of options and offset the risk by using either synthetic or dynamic portfolios.

The Black-Scholes formula has also become one of the key tools used in the pricing of corporate loans (Jarrow and Turnbull, 1995). Corporate debt is harder to price then government debt as corporations can end up in bankruptcy and default on their obligations; this results in corporate and government debt having different term structures. Banks then use a slightly modified form of the Black-Scholes formula in an arbitrage-free model that is based on the spread between the two term structures to calculate the price of bond.

Conclusion

The work of Black, Merton and Scholes has impacted upon numerous fields in the social sciences. Their solving of the PDE has lead to the development of a new field in financial economics and many of the modern day financial engineering techniques are based upon the insights of their original formula. In addition, the practices of short selling and hedging, as well as the introduction of lower cash outlay per transaction has lead to an increase in the efficiency in many financial markets. The 1973 paper is clearly one of the key foundations of modern financial economics.

Bibliography

Arrow, K. 1999. 'In Honour of the Nobel Laureates Robert C. Merton and Myron S. Scholes: A Partial Differential Equation that Changed the World'. *The Journal of Economic Perspectives* 13:4:229–248.

Bachelier, L. 1900. 'Theorie de la Specualtion - PhD Thesis'. Reprinted (1964) in: *The Random Character of Stock Market Prices*. Cambridge MA: MIT Press.

Black, F. 1989. 'How We Came Up with the Option Formula'. *Journal of Portfolio Management* 15:2:4-8.

Black, F. and Scholes, M. 1973. 'The Pricing of Options and Corporate Liabilities'. *The Journal of Political Economy* 81:637–654.

Cuthbertson, K. and Nitzsche, D. 2001. *Financial Engineering: Derivatives and Risk Management*. West Sussex: John Wiley and Sons Limited.
Duffie, D. 1998. 'Black, Merton and Scholes: Their Central Contributions to Economics'. *The Scandinavian Journal of Economics* 100:2:411–423.

Elton, E., Gruber, M., Brown, S. and Goetzmann, W. 2007. *Modern Portfolio Theory and Investment Analysis*. New Jersey: John Wiley and Sons Limited.

Jarrow, R. and Turnbull, A. 1995. 'Pricing Derivatives on Financial Securities Subject to Credit Risk'. *The Journal of Finance* 50:3-129.

Kim, J., Park, G. Y., and Hyun, J. 2007. 'What is the Correct Meaning of Implied Volatility?' *Finance Research Letter* 4:179-185.

Konishi, A. and Dattareya, R. 1996. *The Handbook of Derivative Instruments: Investment Research, Analysis and Portfolio Applications*. Chicago: Irwin Professional Publishing.

Sutton, W. 1990. *The Currency Options Handbook*. Cambridge: Wood-house Faulkner Limited.

THE MAIN ADVANTAGES OF INTERNATIONAL PORTFOLIO DIVERSIFICATION

DAVID MADIGAN

Senior Sophister

The increasingly integrated global financial system has recently been under attack on all sides. This paper by David Madigan emphasizes the benefits of holding an internationally diversified portfolio of assets, and seeks to explain why investors fail to take full advantage of this risk-reducing strategy. He presents evidence suggesting that the phenomenon of 'home bias' has decreased within the Eurozone thanks to the removal of exchange rate risk from the equation. However, he cautions that in times of crisis the correlation coefficient between international financial markets actually tends to increases.

Introduction

In 1974, Bruno Solnik wrote a famous paper that highlighted the merits of international portfolio diversification. Since then, numerous academic papers have been written on the subject outlining the many benefits. One common theme, repeated throughout this body of work, is that by diversifying internationally the overall risk of a portfolio is reduced. This is due to different underlying industrial structures and unsynchronised business cycles between countries across the globe. The first part of this paper highlights the main advantages for international portfolio diversification. The following section of the paper attempts to explain why equity portfolios of investors are typically concentrated in domestic stocks. Finally the recent trend of convergence in worldwide stock market correlations is examined, with a particular focus on the reasons behind increased correlation coefficients during times of financial distress.

The Advantages

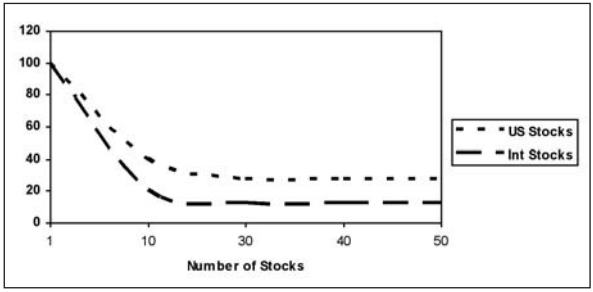
By holding a diversified portfolio an investor reduces risk. The correlation between the returns of the securities that make up a portfolio is critical in determining the associated risk. The total risk of any portfolio is composed of both systematic and unsystematic risk. Traditionally, idiosyncratic, firm-specific (unsystematic) risk can be reduced by holding a diversified portfolio of assets, while market (systematic) risk cannot be diversified away. However, by diversifying internationally, the portfolio's 'beta' (the overall level of systematic risk) is also lowered by exploiting the low correlations of stock market returns across countries. This is due to the fact that the returns on different stock markets are not perfectly positively correlated. The correlation between domestic and foreign securities is lower than purely domestic securities: 'this is due to the monetary, fiscal and industry policies varying from country to country' (Lagoarde-Segot and Lucey, 2007: 2). We expect smaller return correlations between international investments: 'because there are different industrial structures in different countries and because different economies do not follow the same business cycle' (Eiteman et al, 2000: 307). For example, in terms of market capitalization, Ireland's stock market is dominated by financial firms such as AIB and Bank of Ireland. In contrast: 'the German equity market is heavily weighted by cyclical industries, such as automobiles and industry goods, whereas in Switzerland the pharmaceutical companies…have an above average weight' (Freimann, 1998: 33).

Industry betas are different across the board, with the utility sector for the most part having a smaller reaction to market movements, while the computer software industry generally has a larger reaction to market volatility. The recent financial crisis in the UK and the USA has resulted in the Irish stock market suffering on a greater scale in comparison to its EU counterparts. This is due to the domination of the financial firms in the Irish stock market. As such, it is important that investors note the industry structure of countries in their portfolio as two countries will be more highly correlated if their industrial make-up is similar. Some international indices are so industry concentrated that investors are effectively taking a stake in an industry. Return volatility is also related across countries: 'inversely to the number of stocks in the index and positively to the 'Herfindahl' measure of 3 digit concentration within an index' (Roll, 1992: 38). Essentially a country's index is more volatile when it is less well diversified.

Figure 1 shows the reduction in total risk achieved by selecting an equal number of stocks across countries. The benefits of international diversification are clearly substantial. For example, domestic diversification in the United States has systematic risk of 27% while the total risk of a globally diversified portfolio

converges to 11.7% (Solnik, 1974).

Figure 1. International Diversification[1]



It is also evident from the graph that the majority of the benefits (in the form of reduced risk) derived from diversification can be achieved by holding a portfolio of around twenty stocks. Nonetheless, a twenty stock portfolio still has an appreciable level of risk, and moreover the number of stocks to achieve a given level of diversification has increased in recent years (Dimson et al, 2000). It should also be noted that companies can diversify through merging with or acquiring foreign companies. AIB is an excellent example of this, as they have investments in the US, the UK and in mainland Europe. This diversification strategy is beneficial as it may reduce the volatility of profit returns. However if investors can diversify more easily than firms then the benefits of firm diversification are not passed on to investors as they would already have an optimally diversified portfolio.

When comparing the real equity returns across countries in the 20th century, the US fairs particularly well with a real return of 4.32% (Jorion and Goetzman, 1999). It seems that there is no coincidence between this high equity return and the fact that the U.S.A avoided war (on home soil), hyperinflation and political upheaval during this period. This argument is further supported by the fact that Sweden, the second best performing country, also avoided major upheavals: 'This strongly suggests that estimates of equity premiums obtained solely from the U.S market are biased upwards by survivorship' (Jorion and Goetzman, 1999: 955). The cases of Japan and Germany illustrate this point as their returns are distorted by huge losses due to the Second World War. Between 1945 and 1948 Germany equities tumbled by 91%, while Japanese equities lost a whopping 97% in value from 1944 to 1947 (Dimson et al, 2000). However

[1] Source: Solnik (1974)

when comparing the volatility of US returns with the rest of the world between 1920-2000, the volatility of the: 'non-U.S. index is about 10% which is much lower than that of the U.S. market alone, reflecting the fact that the portfolio is spread over a greater number of markets, thus benefiting from imperfect correlations across markets' (Jorion and Goetzman, 2001: 977). This result reflects the benefits of international diversification, which spreads the risk of dramatic events over a large portfolio.

Correlations of equity returns between countries depend on the underlying structural relationship between countries such as colonial past, currency regimes, trade flows, economic growth and industrial make-up. The correlation coefficients of real equity returns among countries worldwide are quite modest, except between countries that are closely linked such as Canada and the United States. In an international study of stock market correlations only 50 out of 276 correlations were above 0.5 and these were mainly between countries that are inextricably linked (Roll, 1992).

Presently, the United States represents about 46% of the value of world stock market in terms of market capitalisation, so one would expect US investors to hold 54% of their portfolio in foreign securities to fully capture the benefits of diversification. However, by the end of 2003: 'US investors held only 14% of their equity portfolio in foreign stocks' (Campbell and Kraussl, 2007: 1239-1240). This is despite recommendations that: 'a US investor, maximising a mean-variance portfolio strategy should hold at least 40% in foreign stocks' (Lewis, 1999 quoted in Campbell and Kraussl, 2007: 1240). Nevertheless there is a steady trend towards greater diversification, with US investors increasing the international weighting in their portfolios from 1% in 1980 to the current weighting of 14% (Dimson et al, 2000). So why do US investors currently hold the vast majority of their portfolios in domestic securities? This phenomenon has been labelled 'home bias' in the financial literature.

Home Bias

Home bias measures the degree to which investors of a given country are: 'overweight in domestic assets and underweight in international assets, as compared to the benchmark portfolio that would weigh home and foreign countries assets according to respective shares in the global financial market' (Fidora et al 2007: 635). Many studies have been undertaken to explain the strong preference of investors towards domestic assets. Exchange rate volatility is a major deterrent for investors as the equity returns can be altered significantly if the exchange rate moves dramatically. For example, the US dollar has

experienced a steady depreciation against the euro over the six month period straddling the last three months of 2007 and the first three months of 2008, with the euro/dollar rate set to hit the unprecedented $1.50 mark in the next few months. Therefore over this time period European investors with large weightings in US equities would have experienced lower real gains than a US domestic investor who had invested in exactly the same stocks. On the other side of the equation, a US investor with a large weighting in European equities would have gained from the favourable swing in the exchange rate. A recent study on this subject states that: 'real exchange rate volatility can explain about 20% of the cross-country variation in equity and bond biases' (Fidora et al, 2007: 633). Derivative instruments such as options and futures can be used to successfully hedge against exchange rate volatility although there is a cost associated with hedging.

Capital controls and the degree of corporate governance have also been highlighted as possible reasons for home bias. An extreme example of capital controls is that a US investor may not have been able to invest in Japan and Germany during the Second World War (Jorion and Goetzman, 1999). Information and transaction costs are also examples of barriers to foreign investment; however, the internet has dramatically reduced the information deficit and consequently reduced information and search costs. Patriotism must also be accounted for when examining investor preferences as investors may feel it is their 'duty' to invest in domestic companies. Remarkably, patriotism is also highlighted as a factor in foreign investment, particularly amongst US investors who exhibit what is known as 'mother country bias' towards countries such as Mexico and Ireland due to the strong ancestral links the US has to theses countries (Dimson et al, 2000). On the other hand, smaller international markets may be perceived as less attractive investments due to their tendency to be: 'less liquid, more prone to price volatility, susceptible to physiological influences and probably less efficient' (Roll, 1988: 33). Cleary there are a diverse range of factors influencing any investment decision.

The European Monetary Union (EMU)

The EMU provides an excellent case study on the reasons for home bias as many factors highlighted in the previous section are eliminated. The introduction of the euro has meant that currency risk disappears completely among participating countries. As a result, the barriers to cross-border investment arising from the cost of hedging are eliminated. Secondly, the common monetary policy has resulted in the convergence of long-term interest rates which has brought about

almost perfectly correlated real risk free rates.

Figure 2. Bond and Equity Bias in the Euro Area[2]

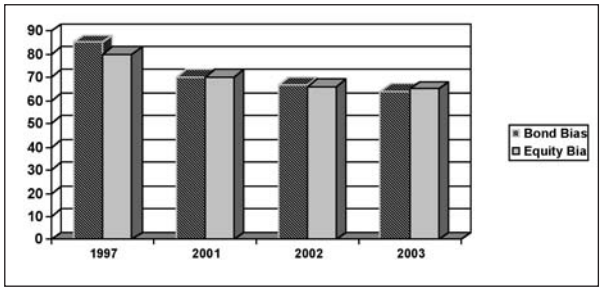


Figure 2 shows how bond and equity bias in the Euro area amongst EMU countries has steadily dropped since the introduction of the single currency. Between 1997 and 2003 equity bias has dropped from 80% to 64% which demonstrates that the reduction in exchange rate volatility has induced EMU investors to increase their holdings of other countries in the monetary union (Fidora et al 2007).

However, the monetary integration has also led to greater trade integration. Due to the increased interconnectedness of the EMU countries, there is now greater business cycle synchronisation. These factors, coupled with the disappearance of currency risk have resulted in stock market correlations increasing for most countries in the EMU as shown in Figure 3.

Figure 3. Return Correlations to an EMU return[3]

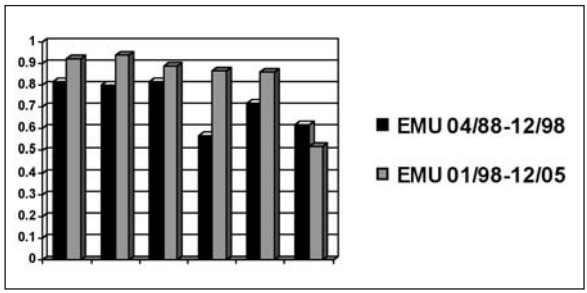


Figure 3 demonstrates that: 'for most countries, correlation coefficients between

[2] Source: Fidora et al. (2007)
[3] Source: Lane and Walti (2006)

the returns of individual EMU participants and the EMU have increased after the introduction of the euro' (Lane and Walti, 2006: 9). The reason for Ireland's low correlation is that it is strongly linked to US performance and to a lesser extent the UK. The resulting higher correlations in the EMU means that is more difficult for a European investor to diversify away risk in the region by holding stocks from other EMU nations.

The Convergence of Worldwide Stock Correlations

Over the past twenty years, the importance of the domestic stock market in many industrialised countries has risen sharply, while at the same time the degree of co-movement amongst international equity markets appears to have increased. Increased financial integration and globalisation are core reasons why correlations between countries have risen. Correlations between Germany, the United States and the United Kingdom have more than doubled from 0.3 to 0.65, whereas correlations between Japan and the same three countries have remained roughly constant at 0.3 (Berber and Jansen, 2005). This trend in correlations has resulted in the weight of the Japanese equities in the optimal world portfolio increasing over time.

However, the general increase in correlations has a downside for investors as: 'national economies are more frequently affected by the disturbances originating in foreign stock markets' (Berber and Jansen, 2005: 833). International stock markets are more highly correlated in periods of high volatility, particularly when it is downward volatility. The correlations between countries may also be higher: 'in some periods of the business cycle, for example periods characterised by high levels of interest rates and dividend yields' (Longin and Solnik, 1995). It has already been highlighted in this paper that the benefits of international portfolio diversification result from the relatively low correlation between country returns so therefore it would be particularly useful to achieve a lower total risk in times of financial market crises. Unfortunately, the empirical evidence suggests that correlation coefficients across countries actually rise during periods of financial distress. Consequently, the benefits from international diversification are reduced just when they are needed most. The Eastern Asia crises of 1997 will now be examined to demonstrate this point.

The Asian Financial Crisis of 1997

Prior to 1997, East Asia had the enviable economic conditions of soaring growth,

low inflation and high saving, with countries such as Thailand, Indonesia, Singapore and South Korea experiencing high growth rates. As a result, the currency crisis that hit East Asia in 1997 was largely unpredicted. The currency crisis started with a devaluation of the Thai Baht on 2nd July 1997 due to massive speculative attacks against it in mid-May 1997. The Baht dropped swiftly, losing half its value relative to the US dollar almost immediately. Due to contagion effects or otherwise, other East Asian countries followed the Thai Baht's plunge and the crisis spread in the form of a string of devaluations and stock market collapses as other East Asian currencies came under speculative attacks. Before the crisis there was almost no co-movement in the stock markets of seven Asian countries: South Korea, Japan, Thailand, Singapore, Indonesia, Hong Kong and Taiwan. However: 'unidirectional and bidirectional linkage among Asian equity markets has increased sharply since the financial stock crisis struck Asia in June 1997' (Jang and Sul, 2002: 103). Most notable is the drastic changes in the co-movement among the four Southeast Asian countries of Hong Kong, Thailand, Indonesia and Singapore. In the pre-crisis period between the 1st of October 1996 and the 31st of May 1997, the correlations of equity returns between Hong Kong and Singapore and between Thailand and Indonesia were 0.35 and 0.1 respectively (Jang and Sul, 2002). From the 1st of June 1997 to the 31st of January 1998 these correlations had jumped dramatically to 0.76 and 0.4 (ibid). This illustrates that the merits of diversification may contract under financial duress.

Conclusion

International portfolio diversification carries significant benefits in terms of lower total risk by taking advantage of imperfect correlation coefficients between countries. This is predominantly due to the fact that stock markets across the world have different industrial structures, which means that their business cycles are also out of synch. An investor from the United States who only diversifies domestically typically holds over two and a half times more risk than an investor who diversifies globally. However, a trend has emerged in the last twenty years whereby global equity correlations have converged. Consequently it is becoming more difficult for investors to diversify risk using this strategy. Nevertheless, equity portfolios of investors are typically concentrated in domestic stocks due to factors such as exchange rate volatility and information asymmetries. The introduction of the euro has demonstrated that the removal of exchange rate volatility reduces home bias in equities and bonds. Finally, the main disadvantage associated with international portfolio diversification is that correlation

coefficients tend to increase when the global financial system is under pressure. This has been highlighted in the case of the East Asian crisis of 1997.

Bibliography

Berber, R.R. and Jansen, W.J. 2005. 'Comovement in International Equity Markets: A Sectoral View'. *Journal of International Monetary and Finance*. 24:832-857.

Campbell, R.A. and Kraussl, R. 2007. 'Revisiting the Home Bias Puzzle: Downside Equity Risk'. *Journal of International Money and Finance* 26:1239-1260.

Dimson, E., Marsh, P. and Staunton, M. 2000. *Triumph of the Optimists, 101 Years of Global Investment Returns.* Oxford: Princeton University Press.

Eiteman, D.K., Stonehill, A.I. and Moffett, M.H. 2000. *Multinational Business Finance*. London: Addison Wesley.

Fidora, M., Fratzscher, M. and Thimann, C. 2007. 'Home Bias in Global Bond and Equity Markets: The Role of Real Exchange Rate Volatility'. *Journal of International Money and Finance* 26:631-655.

Freimann, E. 1998. 'Economic Integration and Country Allocation in Europe'. *Financial Analysts Journal* 54:5:32-41.

Jang, H. and Sul, W. 2002. 'The Asian Financial Crisis and the Co-Movement of Asian Stock Markets'. *Journal of Asian Economics* 13:94-104.

Jorion, P. and Goetzmann, W.N. 1999. 'Global Stock Markets in the Twentieth Century'. Journal of Finance 54:3:953-980.

Lagoarde-Segot, T. and Lucey, B.M. 2007. 'International Portfolio Diversification: Is There a Role for the Middle East and North Africa?' *Journal of Multinational Financial Management* 17:401-416.

Lane, P.R. and Walti, S. 2006. 'The Euro and Financial Integration'. *IIIS Discussion Paper* 139:1-35.

Longin, F. and Solnik, B. 1995. 'Is the Correlation in International Equity Returns Constant: 1960-1990?' *Journal of International Monetary and Finance* 14:1:3-26.

Roll, R. 1988. 'The International Crash of October 1987'. *Financial Analyst Journal* 44:5:19-35.

Roll, R. 1992. 'Industrial Structure and the Comparative Behavior of International Stock Market Indices'. *Journal of Finance* 47:2:3-41.

Solnik, B. 1974. 'Why Not Diversify Internationally Rather Than Domestically'. *Financial Analysts Journal* 30:4:48-54.

HOW SWITCHING COSTS CAN MAKE PREDATORY PRICING RATIONAL: THE ECONOMICS OF WANADOO INTERACTIVE'S PREDATORY STRATEGY

JOHN LAVELLE

Senior Sophister

By setting its prices artificially low, a predatory firm seeks to capture valuable market share, forcing its rivals out of business. Economists have frequently debated the rationality of such a strategy, which aims to recoup the short-term losses by earning supra-normal profits in the long run. John Lavelle analyzes the seminal predation case of 'Wanadoo Interactive' and presents a model seeking to explain their pricing strategy. By emphasizing the role of 'switching costs', specifically in the case of a growing market, he proposes that existing models of predation fail to capture all of the relevant instances in which such behaviour might prove to be rational.

Introduction

Predatory pricing has long been one of the most controversial topics in industrial economics. For decades, many economists disputed the notion that predation – selling at a low cost in order to eliminate rivals from the market – could ever be a rational business strategy. More recently, industrial economists have used game theoretical models to demonstrate that predatory pricing can be rational in certain circumstances, usually when competitors are faced with incomplete information.

This paper argues that modern predatory pricing theory has overlooked an important strategic reason for predation. In the early stages of markets with switching costs, it can be rational for a firm to initially set price below cost in order to attract a large market share at the expense of its competitors. Later, it can raise prices to exploit its large base of 'locked-in' consumers. This strategy is illustrated through a real world instance of predatory pricing, the case of Wanadoo Interactive: the publicly owned French internet provider which was fined € 10.35 million by the European Commission for predatory abuses in 2003.

Part I briefly outlines the economic theory and legal treatment of

predatory pricing. Part II reviews the seminal predation case of Wanadoo Interactive, and argues that conventional theories of predation do not explain why Wanadoo set its prices below cost and incurred large losses between 2001 and 2002. In Part III, there is a discussion of the economics of markets with switching costs. It includes an explanation of how said costs can cause companies to initially set prices low in order to secure a large market share which can be exploited in future. A model of how switching costs induce firms to reduce their prices in order to grow their market share is outlined, and extended to show the conditions under which firms will set price below cost. It is argued that the presence of switching costs is the best explanation of why Wanadoo set its prices at predatory levels. The paper concludes with a brief discussion of the role that economic models of predatory pricing should play in legal proceedings against alleged predators.

I. Predatory Pricing: Economic Theory and Legal Policy

The economic concept of predatory pricing can be broadly defined as a firm foregoing short-run profit with the intention of eliminating a competitor. The firm acts in this manner in order to earn higher profit in the future than it would have earned had the competitor remained in the market.[1] The predatory firm sets prices lower than it would under 'normal' competitive conditions. Its rival cannot afford to compete and leaves the market, allowing the predator to gain market power and set prices above the competitive level. It makes extra long-run profit which compensates it for its short-run losses. Predatory pricing harms the welfare of both consumers, who pay higher prices in the long run, and the firms that are forced out of the market, which lose out on profit.

The spectre of large companies undercutting small enterprises to bankrupt them and monopolise the market has concerned policymakers for over a century. Predatory pricing was outlawed in the United States by the Sherman Antitrust Act of 1890. But, beginning in the 1950s, economic analysis by McGee (1958) and others of the Chicago School cast doubt on the traditional view of predatory pricing. Predation, they claimed, is rarely a rational strategy. Since it is seldom attempted in practice, there is no need for legislation.

More recently, the development of dynamic game theoretical models has allowed economists to identify conditions under which predation may occur (Bolton et al, 2000). Most models of predatory pricing assume some form of

[1] This is loosely based on the definition proposed by Ordover and Willig (1981).

incomplete information. In 'reputation' models, a firm responds to competition by setting low prices in order to establish a reputation as a 'tough' competitor and deter firms from competing with it in other, related markets. In 'cost signalling' models, a predator sets low prices in order to mislead rivals into believing that it has a low cost base and induce their exit from the market. 'Financial market' predation models show how, with imperfect financial markets, firms can respond to competition by setting low prices and making competitors appear unprofitable, preventing them from securing further funding from myopic financers. Finally, in 'test market' predation models, the predator sets artificially low prices to prevent competitors from learning about market conditions and discourage them from remaining in the market. These incomplete information models have somewhat allayed the sceptical attitude to predatory pricing inspired by the Chicago School. There is now, according to Motta (2004), a broad consensus among economists that predation can be a rational strategy.

Due to its negative effect on welfare, predatory pricing is illegal in most jurisdictions. In the European Union[2], predation is forbidden under Article 82 of the Treaty of Rome, which makes it unlawful for a firm to abuse its dominant position by 'directly or indirectly imposing unfair purchase or selling prices'. In order for predatory pricing to be proven, under European case law, it must be established that: first, the alleged predator held a dominant position in the market at the time of the offence; and second, it acted with the intention of eliminating a competitor from the market.

In *AKZO v. Commission* the European courts adopted a cost-based test to establish predatory intent. If a dominant firm sets prices below average variable costs[3] these are presumed to be predatory as: 'there is no conceivable economic purpose other than the elimination of a competitor'. This legal rule of thumb, first proposed by Areeda and Turner (1975), is designed to simplify legal proceedings by providing a clear delineation between normal price competition and 'unnaturally' low prices intended to eliminate rivals.

[2] European Union law is only applicable to cases where predatory pricing 'affects trade between Member States'.

[3] Average variable costs are used as a proxy for marginal costs, which are more difficult to calculate. Prices above average variable cost but below average total cost may be considered predatory if they are found to be part of a plan to eliminate a rival.

II. Case Study: Wanadoo Interactive v. The European Commission

The Wanadoo Interactive case is among the most significant instances of predatory pricing in European case law. On 16 July 2003, the European Commission fined Wanadoo Interactive, a subsidiary of the state-owned France Telecom, for abusing its dominant position in the French high-speed internet market by setting prices at predatory levels (European Commission, 2003). After a series of appeals, on 13 January 2007, the European Court of First Instance upheld the decision to impose a record fine of €10.35 million on the company.

The market in question was the provision of ADSL (a type of high-speed internet access) to the French public. Wanadoo Interactive held a market share of 46% at the beginning of 2002 and was adjudged to have held a dominant position. According to court proceedings, between January 1999 and August 200, Wanadoo set its price for a monthly subscription to its home high-speed internet service below its average variable cost, and between August 2001 and August 2002, below its average total cost. This pricing strategy caused Wanadoo to incur losses until the end of 2002, when the abuse was stopped, and its' expected losses to continue up to 2004. Between January 2001 and September 2002, Wanadoo's market share rose from 46% to 72%, with its competitors' shares shrinking to below 10% each. One rival ADSL provider, Mangoosta, was forced to leave the market. The courts supported the Commission's claim that while holding a dominant position in the market Wanadoo set prices below average variable cost. The legal test for predatory pricing, as established in *AKZO vs. Commission*, was met.

However, while the legal case is relatively straightforward,[4] the economics underlying the Wanadoo Interactive case is less clear cut. Unlike in the United States, European law does not require the competition authority to prove that a company could reasonably expect to recoup losses incurred by pricing below cost. Consequently, in preparing the case the Commission devoted most of its attention to showing that Wanadoo's price was below cost. Significantly less attention was devoted to establishing why it was economically rational for the company to incur such large losses through below cost selling.

On closer examination, none of the standard game theoretical models of predatory pricing can adequately explain Wanadoo's predatory strategy. Most modern theories of predation assume some form of incomplete information but it is debateable whether this is characteristic of the French ADSL market. The

[4] Although France Telecom's appeal disputed both matters of fact and points of law, the Court of First Instance sided with the Commission in virtually all areas of disagreement.

'reputation model' does not seem applicable to the Wanadoo case, as the theory assumes repeated interaction between competitors in a large number of markets – Wanadoo Interactive was not competing with its ADSL competitors in other, more profitable markets, so foregoing profit to establish a reputation as a 'tough' competitor would not have served much purpose. The 'cost signalling' model also seems inappropriate. Wanadoo Interactive was a large, publicly owned company and information about its costs would have been readily available to competitors from published financial accounts. Setting prices low in order to deceive competitors into thinking that it had a cost advantage seems an unlikely strategy. The 'financial market' model of predation is also unlikely to apply. Widespread knowledge about profitability, growth rates and cost levels in the industry would make it difficult for Wanadoo to manipulate market information leading financiers to deny funds to its competitors. Finally, the 'test market' model only attempts to explain predatory abuses in isolated geographical markets, and therefore does not apply in the Wanadoo case.

How then was it profitable for Wanadoo Interactive to pursue a predatory strategy? Based on its analysis of the Wanadoo case and documents recovered from raids on the company's premises, the European Commission (2003) speculated that the below cost pricing strategy was driven by a pursuit of a larger market share:

> 'The abuse on which the Commission has taken action was designed to take the lion's share of a booming market, at the expense of other competitors' (European Commission, 2003).

But can it be a rational strategy to price below cost in order to grow market share at the expense of competitors, as the Commission suggests? If so, how? Textbook theories of predation don't admit such a possibility. The European Commission bases its hypothesis that Wanadoo was attempting to 'pre-empt the market' primarily on documents recovered from raids on the company's premises – not on economic theory.

III. Below Cost Selling in Markets with Switching Cost

Switching costs, market share and pricing strategy

Switching costs are the costs associated with switching consumption from one good to another. Such costs cause two products which are functionally identical to become differentiated after customers have purchased one of them. Klemperer (1995) identifies six types of switching cost:

1. Investment in compatible products: if consumers switch to a new brand of computer, they may also have to invest in new complementary equipment, like a printer or monitor that is compatible with the new brand.

2. Transaction costs: when customers switch their current account to a different bank they must incur costs such as time spent evaluating other banks' products, time spent closing their existing account, the cost of completing documentation necessary to open an account and so on.

3. Learning costs: if a consumer switches to a new brand of computer, he must spend time learning to use the new system.

4. Risk: consumers are uncertain about the quality of untried products and so there is risk attached to switching suppliers.

5. Artificial costs: firms can design products in such a way that customers are penalised financially for switching to another supplier, such as when airlines offer frequent flier programmes.

6. Psychological costs: people's tastes evolve to favour the brand they are using and so there are psychological costs in switching to another, unfamiliar brand.

Switching costs may be borne by producers or consumers – their effect on competition is the same (ibid).

In addition to the six forms of switching costs identified by Klemperer, I

add a seventh: time preference costs. These are due to the empirical observation[5] that people have hyperbolic discount functions – they prefer small gains to larger gains when the small gain is immediate, but when both gains are in the future people prefer the larger of the two. Switching costs, such as the cost of searching for a new supplier, tend to be incurred immediately. In contrast, the benefits of switching to a new product, such as lower prices or improved quality, tend to accrue to the consumer over future periods. If consumers discount hyperbolically, then the impact of switching costs is magnified relative to the benefits of changing supplier.

By making *ex ante* homogeneous products *ex post* differentiated, switching costs make price competition less vigorous and give firms market power over existing customers. Once consumers are 'locked in' to a firm's product by switching costs, the firm can raise prices with reduced fear of losing customers. It follows that the higher a firm's market share, the higher the monopoly profits it can extract from its existing customers. Realising this, firms will set prices low in the early stages of a market's development in order to attain a large share of the market which can subsequently be exploited (Klemperer, 1987a, 1987b, 1995). Under certain conditions, one model predicts that: 'with switching costs, the competition for market share is sufficiently fierce that first-period prices are below firms' costs' (Klemperer, 1987a: 390). In other words, switching costs can make it a rational profit maximising strategy for firms to increase their market share by initially setting price below marginal cost, an *a priori* indication of predatory pricing under European law.

This is an interesting conclusion, with considerable implications for the economic and legal treatment of predatory pricing. It is strange, then, how rarely Klemperer's result is mentioned in industrial organisation literature. In the twenty years since Klemperer's paper, no model has been developed specifically to show the conditions under which switching costs might lead to below cost selling.[6]

---

[5] For a summary of empirical work on hyperbolic discounting, see Angeletos et al. 2007

[6] It is important to distinguish here between predatory pricing; lowering prices in order to force a rival's exit, and the related concept of limit pricing; lowering prices in order to prevent a rival's entry. While there is no model which sets out to show how switching costs can rationalise predatory pricing, Klemperer (1989) does show how switching costs can cause firms to engage in limit pricing.

Klemperer's model of a market with switching costs

The following model is adapted from one of Klemperer's later papers (1995)[7] in order to illustrate how high switching costs can lead firms to set prices low in the early stages of the market before raising them once consumers are locked in. The model is extended to show the conditions under which a firm will set first-period prices below marginal cost.

Consider a two period duopoly, where firms A and B maximise profit by setting prices simultaneously and non-cooperatively in each period. There are N consumers, each with a reservation cost, R, for the products of both firms. Initially, we assume that firms A and B have marginal costs of c_1^A and c_1^B, respectively, in the first period and c_2^A and c_2^B, respectively, in the second period. Both firms discount second period profits by δ, and $\delta \leq 1$. Firms cannot price discriminate.

In the first period, both A and B's products are differentiated[8]. This is represented by placing firms A and B at each end of a line segment $[0, 1]$ at 0 and 1 respectively, with the N consumers evenly distributed along the line segment. A consumer at point y on $[0, 1]$ has a transport cost, Ty, [9] of using A's product and $T(1 - y)$ of using B's product.

In the second period, a fraction of consumers, σ^A and σ^B (where $\sigma^B = 1 - \sigma^A$), have already bought from firms A and B, respectively. If these consumers wish to switch suppliers in the second period, they must pay a switching cost, S. We assume that switching costs are high: $S > R - c_2^A > 0$ and $S > R - c_2^B > 0$. The following assumptions about the dimensions of costs are also necessary:

$$R - 2T > c_1^i \; ; \; R > c_2^i \; , \text{for } i = A, B; \; T > \mid c_t^A - c_t^B \mid, \text{for } t = 1,2$$

Firms set first and second period prices to maximise total profit, V^i, given by:

$$V^i = \Pi_1^i + \delta \Pi_2^i(\sigma^i) \qquad \textit{(for } i = A, B\textit{)} \qquad [1]$$

[7] This model from Klemperer (1995) is more or less a hybrid of the Klemperer (1987a) model referred to above and another of his switching cost models published in the same year (1987b). The result that interests us – that switching costs can, under certain conditions, make it rational for firms to sell below cost – holds in all three cases.

[8] By introducing first-period product differentiation into the price setting model, the lack of realism implied by the Bertrand paradox (where P = c) is avoided. This product differentiation might be caused by different consumers having different learning costs for each product.

[9] The 'transport cost' in the model is simply a way of incorporating product differentiation into the model. T can be thought of a measure of product differentiation and not, literally, the cost of transport.

where second period profit, Π_2^i, depends positively on first-period market share, σ^i. A general solution to this equation is attained by solving for the profit maximising second period price, substituting this into the total profit equation, and then solving for first-period price. As Klemperer (1995) shows, the profit maximising second-period price is:

$$P_2^i = R \quad [2]$$

In other words, firms can charge monopoly prices in the second period as high switching costs prevent existing customers from taking advantage of lower prices elsewhere. This presents a firm with two conflicting motivations when selecting first-period price (given second-period profit). On one hand, it wants to charge a high price to increase first-period profit. On the other hand, it wants to charge a low price in order to capture a larger share of the market for the second period to increase second-period profit. It can be shown that each firm resolves this

$$P_1^A = T + [(2c_1^A + c_1^B)/3] - \delta[R - (2c_2^A + c_2^B)/3] \quad [3]$$

first- and second-period profit trade-off by setting a first-period price:

and symmetrically for P_1^B.

Extension: pricing-below-cost in a market with switching costs

At this point, we depart from Klemperer's analysis and explore the conditions under which a firm will set a first period price below marginal cost. A simplifying assumption is made that firms A and B have symmetric costs[10], so $c_1^A = c_1^B = c_1$ and $c_2^A = c_2^B = c^2$. Therefore [3] is reduced to:

$$P_1^A = T + c_1 - \delta(R - c_2) \quad [4]$$

Rearranging:

$$P_1^A - c_1 = T - \delta(R - c_2) \quad [5]$$

[10] We have already assumed that the difference between firms' costs in each period is small (less than T), so this assumption is not a major departure from the original model.

Under what conditions will first-period mark-up be negative, i.e. price below marginal cost? From equation [5]:

$$P_1^A - c_1 < 0$$
$$\text{if:} \qquad T - \delta(R - c_2) < 0 \qquad [6]$$

Or, rearranging:

$$\delta(R - c_2) > T \qquad [7]$$

In other words, when switching costs are sufficiently large, it is rational for a firm to set first-period price below cost when the discounted value of the mark up consumers are willing to pay in the second period is greater than the 'transport cost', a measure of product differentiation.

Admittedly, the value of this model is mitigated by its simplistic assumptions. A limited time horizon, a fixed number of consumers with identical tastes for the product, no price discrimination, two firms with symmetric costs; all of these are uncharacteristic of the real world. But these assumptions are incorporated primarily for mathematical simplicity and if they were relaxed our conclusion – that switching costs can lead firms to price below cost – may well still hold[11].

Wanadoo Interactive, French ADSL and switching costs

In contrast to standard models of predatory pricing, Klemperer's model of a market with switching costs does appear to strongly resemble the Wanadoo case in several important respects. In the early 2000s, the market for high-speed internet access in France was new and growing quickly – in the seventeen months to September 2002 it grew five-fold. Initially, the market resembled the first-period situation in the game above, with a large number of unattached consumers who could be enticed by a low monthly subscription. Firm A's strategy of pricing below cost in the first period in order to attract market share

[11] For example, another of Klemperer's models (1987b) relaxes the assumption that the number of consumers in the market is fixed, by incorporating a fraction of customers to leave the market after the first period and some new customers to enter in the second period. In this model, as in the example discussed, it can be shown that firms will still set price below cost in the first period for certain parameter values.

seems consistent with Wanadoo's actions: it initially set its monthly subscription below its marginal cost, incurring heavy losses, and its market share increased rapidly during the period of predation.

Like in the hypothetical industry described in Klemperer's model, there are significant costs in switching between high-speed internet providers. As Shapiro and Varian (1999) outline, switching costs are especially prevalent in information industries and ADSL is a case in point. A customer wishing to switch provider would have to spend time and money cancelling their existing subscription. He would have to search for another provider, install the new provider's service and learn to operate the new service. In addition, there may be psychological costs to switching, and the customer faces the risk that the new provider may be of lower quality. Unlike a shopper buying his weekly groceries, households purchasing an internet connection are not confronted with a regular choice about who their supplier will be – their subscription continues until they consciously choose to end it. The presence of continuous subscriptions leads to a strong status quo bias in the industry and means that inducing customers to switch providers is extremely difficult.

Because of the presence of high switching costs, subscribers are to a large degree 'locked in' to their existing internet provider, giving companies the power to charge them a higher price in future periods. The second period in the model, where the cost of switching suppliers is prohibitively high, is a reasonable approximation of the French ADSL market as it developed and growth rates levelled off.

Klemperer's model does not, of course, perfectly describe the Wanadoo case. In particular, the model assumes two symmetric firms – in contrast, Wanadoo was significantly larger than its competitors at the time of the abuse. Motta (2004) asserts that it cannot be rational for a firm which already has a larger market share than its rivals to lower price below cost in order to attract more customers. However, Motta's claim may not be true of a small market that is growing rapidly, like the French high-speed internet market. Intuitively, even if a firm already has a large share in a small but booming market, it may still be more profitable for it to set a low price in order to win over the large number of customers who are about to enter the market, rather than setting a high price and exploiting its small number of existing customers.

While standard models of predatory pricing fail to explain Wanadoo's actions, Klemperer's model of a market with switching costs leads us some way towards a convincing account of why predation was a rational strategy for the company. Wanadoo initially kept its price low and suffered large losses in an attempt to secure a larger customer base at the expense of its weaker competitors. The firm intended to recoup these losses in the future by raising its prices above

its competitors', safe in the knowledge that high switching costs in the ADSL market made customers unlikely to change provider. This narrative broadly corroborates the European Commission's beliefs about the motivation for Wanadoo's strategy, and is consistent with documentary evidence recovered by the Commission from Wanadoo's premises.

Conclusion: The Role of Economic Models of Predation in Legal Actions

Bolton et al (2000) echo the views of many industrial economists when they argue that the current legal approach to predatory pricing is overreliant on simplistic legal rules of thumb. They criticise the below average variable cost test for predatory pricing as deficient as it fails to take account of the motivation behind price-below-cost strategies, and ignores legitimate business justifications for selling below price. They say that competition authorities should support their analyses with economic models of predatory pricing when seeking to prove predatory intent in legal actions. The authors believe that the dynamic game theoretical models of predation developed since the 1970s – reputation, cost signalling, financial market and test market models – identify the main circumstances in which predatory pricing is theoretically rational. Furthermore, these models are capable of explaining companies' behaviour in real world cases of predation.

Existing models of predatory pricing are certainly valuable, and are applicable in some real world instances of predation. But this paper's analysis of the Wanadoo Interactive case shows that textbook predation models do not exhaustively describe the circumstances under which predatory pricing occurs in the real world. Wanadoo's predatory strategy was designed to capture share in a new, growing market with switching costs. Yet there has never been a real attempt to construct an economic model of predatory pricing caused by switching costs. Aside from a small number of general purpose papers on the economics of markets with switching costs – which do not attempt to explain predatory pricing strategies – the existing body of economic theory gives little insight into the Wanadoo case, one of the seminal European predation cases.

Calls for economic models to be given a more central role in proving predatory intent in antitrust cases are, therefore, premature. Before this becomes a plausible option, more robust economic models of predation must be developed to more comprehensively describe the circumstances in which predatory pricing is rational. A model of selling below cost to grow market share, with asymmetric competitors in the early stages of a growing market with switching costs, would make an excellent starting point.

Bibliography

Angeletos, G.M., Laibson, D., Repetto, A., Tobacman, J. and Weinberg, S. 2001. 'The Hyperbolic Consumption Model: Calibration, Simulation, and Empirical Evaluation'. *The Journal of Economic Perspectives* 15:3:47-68.

Areeda, P. and Turner, D.F. 1975. 'Predatory Pricing and Related Practices under Section 2 of the Sherman Act'. *Harvard Law Review* 88:4:697-733.

Bolton, P., Brodley, J.F. and Riordan, M.H. 2000. 'Predatory Pricing: Strategic Theory and Legal Policy'. *Georgetown Law Journal* 88:2239-2330.

European Commission. 2003. Press release: 'High Speed Internet: the Commission imposes a fine on Wanadoo for abuse of a dominant position', 16/07/2003. Viewed at: http://europa.eu/rapid/pressReleasesAction.do?reference=IP/03/1025&format=HTML&aged=1&language=EN&guiLanguage=en

Klemperer, P.D. 1987a. 'Markets with Consumer Switching Costs'. *The Quarterly Journal of Economics* 102:2:375-394.

Klemperer, P.D. 1987b. 'The Competitiveness of Markets with Switching Costs'. *Rand Journal of Economics* 18:138-1 50.

Klemperer, P.D. 1989. 'Price Wars Caused by Switching Costs'. *The Review of Economic Studies* 56:3:405-420.

Klemperer, P.D. 1995. 'Competition when Consumers have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics, and International Trade'. *The Review of Economic Studies* 62:4:515-539.

McGee, J. 1958. 'Predatory Price Cutting: The Standard Oil (NJ) Case.' *Journal of Law and Economics* 1:137-69.

Motta, M. 2004. *Competition Policy: Theory and Practice*. Cambridge: Cambidge University Press.

Ordover, J.A. and Willig, R.D. 1981. 'An Economic Definition of Predation: Pricing and Product Innovation'. *The Yale Law Journal* 91:1:8-53.

Shapiro, C. and Varian, H.R. 1999. *Information Rules: A Strategic Guide to the Network Economy*. Boston: Harvard University Press.

List of Cases Referenced

Case C-62/86 *AKZO Chemie BV v. Commission* .1991.

Case T-340/03 *France Telecom SA v. Commission* (Formerly *Wanadoo Interactive v. Commission*). 30/01/2007.

Case COMP/38.233 Wanadoo Interactive, Commission Decision of 16/07/2003.

NETWORK DYNAMICS, PREFERENTIAL ATTACHMENT AND MARKET LIBERALISATION

ENDA HARGADEN

Senior Sophister

Today's markets are constantly in flux, with firms forced to continually grow and adapt in order to survive in the increasingly interconnected global environment. In this dynamic context, dominant firms frequently emerge holding large shares of their respective markets. By focusing on the preferential attachment of revenue growth to larger firms, Enda Hargaden takes a novel approach to this topic. He expands on existing measures of market concentration, combining these measures with cutting edge network theory. The model he proposes demonstrates significant monopolistic pull, with the result that the initially dominant firm will eventually overpower the other market players. The implications of these findings for future industrial regulation are considerable.

Introduction

The Hirschman-Herfindahl Index (HHI) is a formula for analysing market concentration. This paper borrows from network theory and proposes a model demonstrating that by using the HHI as a benchmark, implementing barriers to entry in growing markets with nonlinear distribution of said growth can be preferable to complete freedom of entry. The advantage of using the HHI over measures such as the concentration ratio is that it gives greater weight to larger firms. It is often used as a proxy measure of competition; however, it fails to take barriers to entry into account, an important consideration for any comprehensive evaluation of competitive market forces within an industry. This paper explores a further problem of simple HHI analysis. It examines a market that displays continuous growth and preferential attachment, demonstrating that such a market will display significant 'monopolistic pull', dependent upon: the initial market shares, the level of growth of the market and the weight of the attachment. These findings are particularly relevant for policy in the case of an industry confronted

by liberalisation from previous monopolisation. Other influential factors discussed include the number of 'footloose' customers and the total number of firms. Interestingly, the results of the model suggest that too many small firms competing in an industry in which one firm has a large share of the market, may well result in a monopoly outcome.

The HHI

The HHI is a standard measure of market concentration used by the EU Commission, the US Department of Justice's Antitrust Division and the Irish Competition Authority, among others. The HHI of an industry is the sum of the squares of its market shares. Formally, where there are n firms competing in the industry and x_i is the percentage market share of firm i:

$$HHI = \sum_{i=1}^{n} x_i^2$$

This returns a value of $0 < HHI \leq 10{,}000$, where perfect competition is represented by a level close to zero[1] while 10,000 implies total perfect monopoly.

It is important to note that the HHI is a measure of market concentration in an industry and is thus indicative of firms' ability to raise prices above their competitive level. The HHI itself does not explicitly consider the observed prices pertaining in the market. An assumption is required that increased market power will have a detrimental impact on consumer welfare. This is common practice; the Irish Competition Authority's official guidelines on mergers state that firms are: 'assumed to pursue maximum profits' (Competition Authority 2002: 5). The assumption that undesirable price increases are ex ante more likely in markets with higher HHIs is retained for this paper.

Network Theory

Network theory primarily concerns itself with the study of graphs which contain points (or nodes) and lines (or edges) connecting the nodes. The number of edges emerging from a particular node represents that node's degree of connectivity.

[1] As at least one firm has a market share >0, the HHI is strictly positive.

For decades, sociologists have used these graphs to illustrate social concepts such as cliques. More recently, physicists and mathematicians have developed network theory for more inferential use and to study networks as stochastic objects with probability densities. Furthermore, the recent past has seen developments in dynamic network theory–the study of networks which develop over time–and it is the concept of dynamic networks which will be utilised in this paper and perhaps which contains the greatest potential for future use within economics.

Network theory is less obscure than the reader may first imagine. Much of the general population are familiar with the 'six-degrees of separation' concept; the hypothesis that that each person can be connected by at most six intermediate acquaintances (or nodes). Although this concept is not strictly true, it provides a useful insight into the 'small world phenomenon'.[2] The network of the world-wide web has been the source of much research. Albert et al. (1999) found that two randomly chosen web pages, out of a possible 800 million, were just nineteen 'clicks' away from each other on average.

Network theory has many real world applications. Preventing the spread of HIV/AIDS can be classed as a problem that requires sufferers to pass the infection onto less than one person on average; that the average infected node creates less than one edge to an uninfected node. Google's system of listing search results, PageRank, is a network-based algorithm whereby pages are ranked in accordance with how many other sites link to them, those links in turn weighted toward sites that are highly linked themselves. The combination of social networking websites' friend-lists and profile views provide invaluable data previously unobtainable regarding the level and distribution of the connectivity of social networks.

The internet has also spawned research into clustering or 'preferential attachment'. Websites become popular based largely on how many sites link to them, and being popular in turn encourages more links. Over time this preferential attachment clusters people to a small number of very popular sites. A similar effect has also been observed in academic citations (Price, 1976). Barabasái and Albert (1999) show that continuously growing networks displaying preferential attachment produce scale-free networks that are dominated by few nodes of very high relative degree. If we consider a market as a network where consumers are connected via purchases (edges) to firms (nodes),

[2] Research from the University of Virginia has shown that assuming an edge is created whenever two actors appear in the same production; the average number of actors (nodes) between Kevin Bacon and 250,000 other actors studied was just 2.96. See http://oracleofbacon.org/ for details.

the Barabasái and Albert result can become extremely important. The next section models a market in such a way.

The Model

Begin with a vector $R_t^n = (r_t^i, \ldots, r_t^n)$ of the revenues of firms *1* through to *n* at time *t*. By defining total market revenue:

$$\overline{R}_t = \sum_{j=1}^{n} r_t^j$$

We can create a vector of market shares by dividing each firm's revenue by total market revenue:

$$x_t^i = \frac{r_t^i}{\overline{R}_t}$$

We now allow for constant growth in market revenue at rate γ. The crux of the model is that this growth in revenue γ is distributed disproportionately; there exists a preferential attachment of it to larger firms. This is not incompatible with standard economic analysis. The HHI itself applies exponential additional importance to higher market shares. The nonlinear distribution is at a weight w[3]. Specifically:

$$r_t^i = r_{t-1}^i + \frac{(x_{t-1}^i)^\omega}{\sum_{j=1}^{n}(x_{t-1}^j)^\omega}\gamma\overline{R}_{t-1}$$

$$x_t^i = \frac{r_{t-1}^i}{\overline{R}_t} + \frac{(x_{t-1}^i)^\omega}{\sum_{j=1}^{n}(x_{t-1}^j)^\omega}\gamma\frac{\overline{R}_{t-1}}{\overline{R}_t}$$

[3] w could take several forms. For this model, w should be considered a scalar whose magnitude is >1 and that w is constant for all firms at all time periods. The condition that w>1 is necessary for the clustering effect of this model to occur. As w represents the rate at which new revenue tends to larger firms, an extension of this essay could be to include the price of firm i's product as a determinant of w_i.

But: $$r^i_{t-1} = x^i_{t-1}\overline{R}_{t-1}$$

and $$\frac{\overline{R}_{t-1}}{\overline{R}_t} = \frac{1}{1+\gamma}$$

$$x^i_t = x^i_{t-1}\frac{1}{1+\gamma} + \frac{(x^i_{t-1})^\omega}{\sum_{j=1}^n (x^j_{t-1})^\omega}\frac{\gamma}{1+\gamma}$$

Defining: $$\chi^i_t \equiv \frac{(x^i_{t-1})^\omega}{\sum_{j=1}^n (x^j_{t-1})^\omega}$$

Therefore: $$x^i_t = \frac{1}{1+\gamma}[x^i_{t-1} + \gamma\chi^i_t]$$

Firm i's market share depends positively on the exogenous growth rate of revenue γ, last period's market share, and the impact of preferential attachment. Crucially, the impact of preferential attachment depends not just on a firm's market share but also their market share relative to the market share of other firms in the industry. To investigate if an equilibrium exists, we set:

$$\Delta x^i_t = 0$$

$$\Delta x^i_t = \frac{\gamma}{1+\gamma}[\chi^i_{t-1} - x^i_{t-1}]$$

$$= 0 \iff \chi^i_{t-1} - x^i_{t-1}$$

$$\iff \frac{(x^i_{t-1})^\omega}{\sum_{j=1}^n (x^j_{t-1})^\omega} = x^i_{t-1}$$

$$\iff (x_{t-1}^i)^\omega = x_{t-1}^i \sum_{j=1}^{n} (x_{t-1}^j)^\omega$$

$$\iff (x_{t-1}^i)^{\omega-1} = \sum_{j=1}^{n} (x_{t-1}^j)^\omega$$

$$\iff x_{t-1} = e_j$$

$$(x_{t-1})^T = [1, 0, 0 \ldots 0]$$

Steady-state is observed when one firm has total market share, i.e. when complete monopoly is observed. The practical implications of this result will be discussed in the policy considerations section.

Numerical Example

To provide a more illustrative example, allow γ = 0.2, w = 2 and T = [60 10 10 10 10]. In this case, the market shares evolve over time such that:

Time	Firm 1	Firm 2	Firm 3	Firm 4	Firm 5	HHI
1	0.6	0.1	0.1	0.1	0.1	4000
2	0.65	0.0875	0.0875	0.0875	0.0875	4351
3	0.697	0.0757	0.0757	0.0757	0.0757	5088
4	0.74	0.0649	0.0649	0.0649	0.0649	5648
5	0.7783	0.0554	0.0554	0.0554	0.0554	6181
10	0.9055	0.0236	0.0236	0.0236	0.0236	8222
15	0.9613	0.0097	0.0097	0.0097	0.0097	9244
20	0.9843	0.0039	0.0039	0.0039	0.0039	9689

This simple numerical simulation provides a striking example of the monopolistic pull the model presents. Revenue growth of 20% is not extraordinary and the preferential attachment of w=2 is substantially lower than

power laws observed in some real world scenarios. Nonetheless after 5 periods of this model's evolution, the dominant player's market share increased from 60% to 77.83% and the HHI increased by 54.5%. By the tenth period, the dominant player's market share exceeded 90% and the HHI had more than doubled to over 8000.

Policy Considerations

The model demonstrates that if a market displays continuous growth and preferential attachment–a well-documented occurrence – it will also display significant 'monopolistic pull' dependent upon the initial market shares, the level of growth of the market and the weight of the attachment. This is particularly relevant for industries that face liberalisation from previous monopolisation. It is regularly observed that previous monopolists maintain dominant market shares: ComReg (2007: 12) report that nine years after telecommunications deregulation, Eircom's market share was still 69%.

The initial market share of the dominant firm and consequently the number of firms a market can bear in this context is ultimately related to the 'footloose' fraction of consumers, ρF who are inclined to move from the dominant firm.[4] By definition, new entrants must compete for these customers among themselves and so average market share falls as the number of firms increase. As already noted, this increases the effect of preferential attachment in our model.

Specifically, regulators wishing to avoid clustering must choose n firms in the market such that:

$$\frac{\rho F}{n} = 1 - \rho F$$

$$n = \frac{\rho F}{1 - \rho F}$$

$n > 0$, so no desirable free market solution exists if $\rho F < 0.5$. Values of ρF observed in real life suggest that n should rarely exceed 10. In this model, $n > 10$ would portray an example the oft-cited 'race to the bottom' effect, where

[4] Editor's note: The 'footloose' fraction of customers may well be dependent upon the level of 'switching costs', which can 'lock' customers into a particular firm or product. This concept is explored in the next essay on predatory pricing. Interestingly, these costs can be of a psychological nature.

additional competition hastens a monopolistic outcome.

Under the preferential attachment distribution of new revenue, ultimately the largest firm will conquer the market. In this regard, the model is intended to be illustrative rather than precise. The lower the dominant firm's initial market share, the weaker the monopolistic pull. Furthermore, the longer it takes to conquer a market, the longer rival firms have to respond and the longer competition authorities have to attempt to rectify undesirable outcomes. Thus the initial liberalisation of and specifically the number of firms allowed to enter the market can be vital to the market outcome and these outcomes may not be consistent with those predicted by Contestable Market Theory, among others. The model points to the use of alternative post-liberalisation policies such attempting to lower w or increasing ρF.

Conclusion

This essay sought to integrate the findings of the field of scale-free networks into the context of economic regulation. The model presented showed that a growing market with preferential attachment displays significant monopolistic pull. The results and conclusions deriving from the model conflict somewhat with those of contestable market theory, or at least question the use of simplistic static HHI analysis for market liberalisation policies. It is interesting to consider how future advances in the study of networks may provide insights into and help shape regulatory policy in the future. It is hoped this essay will encourage some future application of network theory into economics.

Bibliography

Albert, R., Jeong, H. and Barabási, A.L. 1999. *The Diameter of the World Wide Web*. Viewed at: http://xxx.lanl.gov/PS_cache/cond-mat/pdf/9907/9907038v2.pdf

Barabási, A.L. and Albert, R. 1999. 'Emergence of Scaling in Random Networks'. *Science* 286:509-512.

The Competition Authority 2004. 'Notice In Respect of Guidelines for Merger Analysis'. Viewed at: http://www.tca.ie/controls/getimage.ashx?image_id=114

Commission for Communications Regulation. 2007. *Irish Communications Market: Quarterly Key Data – December 2007*.
Viewed at: http://www.comreg.ie/_fileupload/publications/ComReg07106.pdf

Demsetz, H. 1968. 'Why Regulate Utilities?' *Journal of Law and Economics* 11:1:55-65.

Newman, M.E.J. 2001. 'Clustering and preferential attachment in growing markets'. *Physical Review* E 64:025102.
Newman, M.E.J., Barabási, A.L. and Albert, R. (eds.) 2006. *The Structure and Dynamics of Networks*. Oxford: Princeton University Press.

Price, D.S. 1976. 'A General Theory of Bibliometric and Other Cumulative Advantage Processes'. *Journal of the American Society for Information Science* 27:5-6:292-306.

FORECLOSURE: A CASE STUDY FROM THE PHARMACEUTICAL INDUSTRY

CONOR FLANAGAN

Senior Sophister

The patented protection of intellectual property rights is fundamental if the incentive to innovate is to be maintained. In this paper, Conor Flanagan examines the contentious issue of foreclosure in the pharmaceutical industry. Genzyme, as the sole providers of an essential medical treatment, held considerable power with regards to price setting. Although Conor recognizes that the firm in question foreclosed on the downstream market, this was acceptable given their right to extract monopoly profit from a patented product. By distinguishing between static and dynamic efficiency, he concludes that the firm in question was incorrectly convicted, to the detriment of societal welfare.

Introduction

This essay analyses a supposed foreclosure in the pharmaceutical industry and its subsequent investigation by the United Kingdom's Office of Fair Trading (OFT). Section I reviews the terms and concepts that are essential to understanding the case. Section II gives the background to the case, defines the relevant markets, assesses market power and finds that foreclosure did take place. Section III argues that there was no obvious anti-competitive motive for the foreclosure. Section IV sifts through the welfare implications of the foreclosure and the OFT's investigation. While static efficiency may have been harmed, the benefits to dynamic efficiency mean the foreclosure is justifiable in welfare terms. The OFT's erroneous finding of abuse may itself have caused significant harm. The essay concludes with a discussion of the broader implications of the case.

I. Foreclosure and Margin Squeeze

> 'Foreclosure refers to a dominant firm's denial of proper access to an essential good it produces, with the intent of extending monopoly power from that segment of the market to an adjacent segment' (Rey and Tirole, 2007: 2148).

There are a number of exclusionary strategies a firm can use to achieve foreclosure such as bundling, tying, refusal to supply and exclusive distribution. A practice known as a margin squeeze is of interest in this case:

> 'A price [or margin] squeeze arises when a [firm], with market power in the provision of an 'essential' upstream input prices it, and/or its downstream product or service in such a way and for a sufficiently long period to deny an equally or more efficient downstream rival a sufficient profit to remain in the market' (Crocioni and Veljanovski, 2003: 30).

There are two criteria which must be satisfied in order to judge that a margin squeeze has occurred. Firstly, it should not be of a temporary or short-term nature but of long enough duration to have an exclusionary effect. Secondly, it should have the effect of making equally or more efficient downstream competitors unprofitable (Crocioni and Veljanovski, 2003).

An imputation test can determine if the second criterion is met. The EU employs two versions of the test to determine if a margin squeeze has taken place:

> 'The dominant company's own downstream operations could not trade profitably on the basis of the upstream price charged to its competitors by the upstream operating arm of the dominant company' (EU Commission, 1998 quoted from Crocioni and Veljanovski, 2003: 50).

> 'The margin between the price charged to competitors on the downstream market for access and the price which the [upstream firm] charges in the downstream market is insufficient to allow a reasonably efficient service provider in the downstream market to obtain a normal profit' (ibid: 50).

II. Case Study: Genzyme and the Market for Enzyme Replacement Therapy[1]

Enzyme replacement therapy (ERT) is the preferred treatment for Type I Gaucher Disease[2], a rare enzyme disorder affecting approximately 180 people in the U.K. Cerezyme, produced by Genzyme, is the only ERT drug on the market.

There are two relevant product markets in this case.[3] The first is the upstream market for drugs to treat Gaucher Disease. Other methods can be used to treat Gaucher Disease, but they are relatively ineffective and are authorised to be used only when ERT is unsuitable. Genzyme faces no current competition in the upstream market on the demand side. The entrance of new suppliers, either now, or in the short- to medium term, is not particularly likely. Genzyme holds patents for the substance itself, and over the production process. Furthermore, the production methods and knowledge that other biotechnology and pharmaceutical firms have developed are not easily transferable. In addition to this, potential entrants will not be incentivised to enter, as Genzyme was. They are unlikely to be afforded the benefits of an 'Orphan Medicinal Product'[4] classification now that a treatment exists. With no current competitors and significant barriers to entry Genzyme's degree of dominance is such that it could be classed as a monopolist in the upstream market.

The second relevant market is the downstream market for home treatment.[5] There are already a number of firms that supply a similar style of home treatment to patients suffering from other illnesses. The conversion costs would not be so high as to block entry. Cerezyme is an essential facility for firms wishing to take part in this market. A firm that could not access Cerezyme could not offer a service. Once firms have access to Cerezyme, the downstream market offers some scope for competition.

The NHS operates and makes agreements on a national basis. The patients

---

[1] This section draws heavily on the Decision of Director General of Fair Trading (27/03/2003) 'Exclusionary Behaviour by Genzyme Limited', Office of Fair Trading, No. CA98/3/03, hereafter OFT (2007)

[2] There are two other, more serious, types of Gaucher disease for which there is no treatment. From this point on ,Type I Gaucher Disease is referred to as Gaucher Disease

[3] To allow a structured and clearly bounded analysis of market power, potential abuses and welfare effects, it is necessary to determine the relevant markets, from both a geographic and product market point of view.

[4] See Section IV for further discussion of this classification.

[5] Home treatment involves provision by a nurse, refrigerated storage and delivery of Cerezyme.

are U.K. residents receiving services at home. This is sufficient to define the United Kingdom as the relevant geographic market.

Following a complaint by Healthcare at Home (HH), the OFT launched an investigation into Genzyme's behaviour. Genzyme supplies the NHS with Cerezyme at a price regulated by the Pharmaceutical Price Regulation Scheme. The NHS pays the same price for the drug whether it carries out the treatment themselves, or if Genzyme, or one of their agents carries out the treatment at a patient's home. Home treatment is, in effect, bundled with the drug at no additional cost.

HH provided home treatment on behalf of Genzyme until 2003, when the contract that it had successfully bid on expired. It was paid a fixed fee by Genzyme. When the contract ended, Genzyme appointed one of its subsidiaries to provide home services instead. HH wanted to stay in the market and so tried to buy Cerezyme from Genzyme. Genzyme charged HH the same price that it charged the NHS for Cerezyme and home treatment, leaving HH unable to cover costs, let alone make a profit.

This pricing policy, which prevented HH from effectively competing with Genzyme in the downstream market, would undoubtedly lead to Genzyme 'failing' both imputation tests discussed above. The price that Genzyme charged HH for Cerezyme, combined with Genzyme's bundling of home treatment with Cerezyme to the NHS, ensured no firm buying Cerezyme at that price could be profitable. Genzyme also fulfilled the first criteria for a margin squeeze, as there was nothing to indicate that their pricing policy was of a transitory nature. Genzyme did indeed employ a margin squeeze, by means of bundling, to foreclose on the downstream market. The OFT investigation found this to be an abuse of market power.

III. Motives for Foreclosure: Did Genzyme have Anti-Competitive Intentions?

It has been established that Genzyme foreclosed on the downstream market. What has not been explained is why? Did Genzyme have an anti-competitive motive or intent? The traditional view, that foreclosure is motivated by a desire to extend monopoly power, was to some extent rebuked in the 1970s by the Chicago School (Crocioni and Veljanovski, 2003; Rey and Tirole; 2007). They argued that since there is only one final product, there is only one monopoly profit. It does not pay to extend a monopoly. Any extra profit extracted downstream will be at the expense of profit upstream. Foreclosure is benign and pro-competitive. A monopolist could only profitably foreclose on a market if it was more efficient

that anyone else in the market.

Later work rightfully asserted that the results of the Chicago model are contingent on two assumptions (Crocioni and Veljanovski, 2003). The first assumption is strict complementarity of the upstream and downstream inputs.[6] The second is a competitive downstream market. Anti-competitive foreclosure could be rational if these assumptions were relaxed. Crocioni (2007) argued that the assumptions behind the Chicago model do hold in the case of Genzyme. A fixed amount of Cerezyme is needed to provide a fixed amount of home treatment. The downstream market is, as discussed above, competitive, or at least has the potential to be so.

Crocioni (2007) also discussed the concept of 'dynamic' foreclosure. A firm may engage in 'dynamic' foreclosure not to extend its monopoly downstream, but to prevent firms from eventually threatening its upstream monopoly. It is hard to see how Genzyme's foreclosure could be of a 'dynamic' nature. The skills and knowledge developed by firms in the downstream industry are not the same skills and knowledge that would be needed to overcome the upstream entry barriers.

The OFT developed a more complex argument derived from this concept of 'dynamic' foreclosure. It argued Genzyme could increase the upstream barriers to entry by foreclosing on the downstream market. Potential entrants into the upstream market need authorisation to sell their product. To get this authorisation, entrants need to test their product on patients. Genzyme, with its control of the downstream market, controls access to patients. They could prevent potential entrants from testing their products and thus entering the market by refusing to supply patients with any drugs but their own. What the OFT overlooked was that these patients need not be tested in the U.K. Authorisation is a European Union matter. Genzyme does not operate in the downstream market, and hence does not have the ability to control access to patients in every country where it sells Cerezyme. It would be possible for a firm to use the results of tests in a different jurisdiction to gain authorisation to compete in the U.K. market.

Rey and Tirole (2007) present a model in which foreclosure is motivated by a wish to retain, rather than extend, the upstream monopoly. Unless all transactions are fully observable, an upstream monopolist might not be able to credibly commit to supplying each downstream buyer with a quantity sufficient to produce the monopoly output. Although the upstream firm has monopoly power, it cannot achieve a monopoly outcome. If the upstream monopoly were

[6] This means the proportion of upstream inputs needed to produce the downstream output is fixed.

efficiency enhancing, then this type of foreclosure would be pro-competitive.

The upstream monopolist can solve the commitment problem by excluding all but one firm from the downstream market. This can be done by integrating downwards and refusing to supply other firms. An alternative way to solve the commitment problem is to deal exclusively with a single downstream firm. Genzyme's actions were in keeping with the behaviour of a firm trying to solve the commitment problem. Prior to the margin squeeze strategy, Genzyme dealt exclusively with HH. Once it is recognised that a margin squeeze and a refusal to supply are different means to the same end (Crocioni, 2007), it becomes clear that the downstream market was always foreclosed. All that changed was the method of foreclosure.

Why the switch from exclusive distribution to vertical integration? One can only speculate. Perhaps Genzyme learned over the course of the contract that the size of the market combined with the transaction, contracting and potential reputation costs meant that compared to exclusive distribution, vertical integration was a preferable method of foreclosing the market. The onus is on the competition authority to prove guilt, not the accused to establish innocence. In this case there was nothing to suggest that Genzyme's foreclosure had an anti-competitive motive.

IV. Welfare and Efficiency: Consequences of Foreclosure

Foreclosure is problematic from a welfare standpoint because it eliminates downstream competition. Competition's value is not intrinsic, but rooted in its' ability to improve efficiency and welfare. There are three types of efficiency to be considered: allocative, productive, and dynamic.

The chief problem associated with a monopoly is its allocative inefficiency. With prices above, and quantity below their competitive level, there is a deadweight loss. However, if demand is perfectly inelastic, the demand curve is vertical and there is no dead weight loss (Motta, 2004). Furthermore, the less elastic demand is, the smaller the deadweight loss would be. In the Genzyme case, it is reasonable to assume downstream market demand to be extremely inelastic. The drug and home treatment are complementary goods and the latter is by far the cheaper of these. When two goods are complementary, the cheaper of the two is more inelastic. The greater the price differential between the complements, the greater is the difference in elasticities. This downstream inelasticity is compounded by the upstream inelasticity of Cerezyme. Pharmaceuticals, such as Cerezyme are, by their very nature, highly inelastic. Health insurance, particularly government administered insurance such as in the

U.K. market, decreases the incentive for agents to consider prices, further adding to inelasticity. Consequently the loss of allocative efficiency resulting from the foreclosure would be very small.

The incentives of society and the firm are aligned with regard to productive inefficiency. A common solution to productive inefficiency is to outsource onto a competitive market. If productive inefficiency were a major concern for Genzyme, it would not have integrated vertically, as it did. If productive inefficiency was not a major worry amongst the informed decision makers within Genzyme, then it should not be a major concern for society either.

Although the effect on static efficiency may be low, the foreclosure does lead to a transfer of surplus from consumers to Genzyme. This may be undesirable from a static consumer surplus perspective. However, it is desirable from a dynamic efficiency standpoint. Dynamic efficiency is concerned with innovation. Patents allow those who innovate a temporary monopoly over the fruits of their research. Without them, firms would be less willing to take on the risks, investment and fixed costs associated with innovation. Others could free ride on their research, which is non-excludable, driving down profits and leaving the innovating firm unable to recover their fixed costs. A balance has to be struck between fostering competition and creating the incentives for innovation. In trying to solve this trade-off, there are five circumstances in which a competition authority should give more to weight static than dynamic considerations (Crocioni, 2007). In the Genzyme case, two of these circumstances are of particular relevance: first, if the downstream market is large compared to the upstream market; second, if imposing an obligation to supply is relatively easy, low cost and unlikely to cause distortions.

Setting the price at which Cerezyme would be available to downstream firms would be complicated and costly when compared to the size of the downstream market, which is small, both relative to the upstream market, and in absolute terms. The circumstances of this case imply the OFT should have weighed the dynamic considerations much more heavily than static considerations, which were small to begin with.

Even from a legal standpoint, refusing to supply intellectual property rights is an abuse only in exceptional circumstances, such as in Magill (Whish, 2003). In this case, it was ruled that television broadcasters were obligated to supply a magazine, Magill, with the right to print T.V. listings. The circumstances were said to be exceptional because the innovation needed to develop T.V. listings was not worthy of legal protection (Motta, 2004). These circumstances certainly do not apply in the Genzyme case, because the innovation in question is undoubtedly worthy of legal protection.

From both an economic and legal perspective, there was nothing

illegitimate about Genzyme extracting a regulated monopoly profit. The patent endowed Genzyme with this right. The OFT was wrong to consider the 'substantial profits' resulting from the foreclosure to be proof of an abuse of market power. The very purpose of a patent is to ensure an innovator is rewarded for their innovation. Foreclosure is not intrinsically harmful. It can improve dynamic efficiency by: 'compensating the bottleneck for its investment or activity' (Rey and Tirole, 2007: 2201).

Cerezyme would never have been developed without the incentive of a patent. By taking away the benefits of this patent ex post two hands of government, working independently, held up Genzyme. This may benefit consumers in the short run, but ultimately will hurt dynamic efficiency not just in this market, but also in the other sectors of the economy where innovation and patents are important. The harm that a competition authority can do to dynamic efficiency by committing a type 1 error, a 'false positive', is far greater in an emerging market,[7] such as the upstream market, where innovation is important. Although the government did not make a 'conscious' decision to hold up Genzyme, it did show an inability to live up its obligations and enforce the patent. In future this will lessen the government's ability to credibly commit to protecting patents, deterring firms from investing in the risky process of innovation. It is imperative for competition authorities to tread lightly when dealing with cases in emerging markets.

Cerezyme was classified as an 'Orphan Medicinal Product' by the E.U.[8] This granted Genzyme benefits above and beyond that of a normal patent[9]. By wrongly punishing the producer of an 'orphan' drug the OFT discouraged innovation where it was most needed, in an area already lacking in incentives. It is a telling indictment of the OFT's investigation that its report makes no reference to consumer harm when discussing the supposed abuse and the penalty to be imposed.

When considering the welfare implications of this case, it is certainly relevant to consider the costs of the investigation. The market for Cerezyme is small and the market for downstream provision smaller still. Even if there were efficiency gains to be made, they would have to be quite large to overcome the costs to society of the investigation, and the legal proceedings that followed.

Although there is scope for downstream competition, how desirable is it

---

[7] Crocioni (2007) discusses both emerging markets and the cost of errors by competition authorities when dealing with them, in great detail.

[8] This classification is designed to encourage innovation in the treatment of rare diseases, where a low potential user base inherently limits profits.

[9] Such as extended marketing exclusivity, research grants and fee exemptions

really? Markets do not function without costs. They have transaction and contracting costs. These costs can be significant when a market is as small as the downstream market in this case. The NHS was opposed to downstream competition. It opposed Genzyme's strategy, not because it prevented competition, but because it wanted a single provider of home treatment for all drugs (OFT, 2007). Genzyme itself obviously did not want downstream competition either. Who did apart from HH?

A final potential welfare concern is that assets were stranded with HH as a result of the foreclosure. However, it should be remembered that HH did willingly take on this risk when it signed the contract with Genzyme. The competition authority should not provide ex post insurance for a firm after a risky contract it signs does not go in their favour. This will only encourage inefficient decision-making and moral hazard. Furthermore, there may be a potential for the assets to be converted for use in providing home treatment for other drugs.

Conclusion

Genzyme foreclosed on the downstream market with the intention of re-capturing its monopoly profit. It did this first via dealing exclusively with HH, and then by vertically integrating and initiating a margin squeeze. It is clear the OFT's finding of abuse was wrong and harmful to both total welfare and long-run consumer surplus, once dynamic welfare is considered.

This case demonstrates foreclosure can have positive effects.[10] Competition authorities should treat foreclosure with a rule of reason, not per se rules. Different means (exclusive distribution, bundling, and margin squeeze) to the same end (foreclosure) should be treated consistently. When nothing but the method of foreclosure changes, intervention by a competition authority is unwarranted.

Competition authorities should also be cautious when using imputation tests. They are a poor substitute for a test that considers the change in consumer surplus or welfare resulting from a foreclosure. Although they may reveal the existence of a margin squeeze, and thus foreclosure, they say little about the possible benefits of the foreclosure. Imputation tests do tests have

[10] As well as serving to reward firms for innovating and snuffing out markets whose existence is not welfare enhancing, it also can reduce monitoring costs, uncertainty relating to associating a product with a poor quality downstream service/good and other costs associated with a competitive market (Rey and Tirole, 2007).

value, but they are no *deus ex machina*.

Competition authorities should be careful when dealing with intellectual property rights, especially in emerging markets. Errors can cause great harm beyond that of the market in question. Property rights are the natural solution to expropriation and a fundamental right (Motta, 2004). Competition authorities should be wary of revoking them. To do so creates uncertainty for firms considering investing in innovation. Rules need to be developed regarding the obligation to supply essential facilities, which result from innovation, and are under patent. By removing ex post uncertainty, such rules would encourage innovation and improve welfare.

Bibliography

Crocioni, P. 2007. `Leveraging of Market Power in Emerging Markets: A review of cases, literature, and a suggested framework'. *Journal of Competition Law and Economics* 1-82 doi:10.1093/joclec/nhm018.

Crocioni, P. and Veljanovski, C. 2003. `Price Squeezes, Foreclosure and Competition Law'. *Journal of Network Industries* 4:28-60.

Decision of Director General of Fair Trading. 2003. 'Exclusionary Behaviour by Genzyme Limited'. 27/03/2003. *Office of Fair Trading* No. CA98/3/03.

EU Commission. 1998. 'Notice on the application of competition rules to access agreements in the telecommunications sector – Framework, Relevant Markets and Principles'. Official Journal C265:2-28.
Viewed at: http://ec.europa.eu/comm/competition/liberalization/legislation

Motta, M. 2004. *Competition Policy, Theory and Practice*. New York: Cambridge University Press.

Rey, T. and Tirole, J. 2007. 'A Primer on Foreclosure' in Armstrong, M. and Porter R. (eds.) *Handbook of Industrial Organisation, Volume 3*. Amsterdam: North-Holland.

Whish, R. 2003. *Competition Law*. London: Lexis Nexis.

MONEY ILLUSION – A REAL PHENOMENON?

CHARLIE NOLAN

Senior Sophister

The price mechanism is a fundamental cog in the self-equilibrating market machine of classical economic theory. Unfortunately, prices are not quite as flexible as the classics maintained. 'Sticky' prices, as they are euphemistically labelled, are a symptom of a much more serious ailment: money illusion. The ability of inflation to erode nominal worth is frequently misunderstood, and more generally, nominal values are erroneously interpreted as being real by supposedly rational economic agents. Charlie Nolan investigates the existence of this irrational anomaly. By considering a number of psychological biases and presenting a review of the empirical evidence, he concludes that the illusion is very real.

Introduction

> 'We have standardised every other unit in commerce except for the most important and most universal of all, the unit of purchasing power. What businessman would consent for a moment to make a contract in terms of yards of cloth, or tonnes of coal? And leave the size of the yard or tonne to chance? We have standardised even now the unit of electricity, the Ohm, the Kilowatt, the Ampere and the Volt. But the dollar, the most important unit of all, is still left to the chances of gold mining' (Fisher, 1913: 214).

Money illusion refers to the confusion of nominal values with real ones: 'An individual will be said to be suffering from money illusion if his excess demand functions for commodities do not solely depend on relative prices and real wealth' (Patinkin, 1965: 2). This paper will begin by examining money illusion and some other closely related psychological biases. By considering the empirical research undertaken by different sets of monetary economists, it is suggested that money illusion is very much a real phenomenon. There is a discussion of how

supposedly rational economic agents can be so easily confused, and why, in many ways, nominal values have become the bedrock of 'popular' economics. Some specific aspects of money illusion are explored, with special regard to the stock market and the housing market. Finally, the effects of money illusion on labour markets are reviewed.

Money Illusion and Related Psychological Biases

Money illusion is closely related to a number of other psychological biases. In a perfect world in which only real values matter, money is a veil and nominal prices have no bearing on decision-making. Life is not, however, based on textbooks, and some of the aforementioned psychological biases prevent the individual from piercing this 'veil' (Brunnermeier and Julliard, 2006). Life is dynamic, and irrationality may certainly exist in the short run.

The 'framing effect' states that alternative representations (framing) of the same decision problem can lead to substantially different behaviour (Shafir et al., 1997). The authors maintain that, to a large extent, an agent's performance depends significantly on whether a problem is phrased in nominal or real terms. Assuming that a problem is phrased in nominal terms, the agents will prefer the nominally less risky option to the other option, which is more risky in real terms. That is, they avoid nominal risk rather than real risk. The reverse is also true; a situation phrased in real terms will likely encourage the agent to avoid real risk over nominal.

These assertions led to empirical research in cognitive psychology which indicated that 'alternative representations of the same situation can lead to systematically different responses' (ibid: 343). Consider a person who receives a 2% rise in nominal wages in times of 4% inflation. Note that we assume this person is aware of inflation and momentarily ignore other factors such as the possible social significance of the salary rise. Naturally this person would be happier with the same nominal rise in times of no inflation. However, because this nominal change is positive, we expect him to be happier than with a 2% nominal wage cut in times of no inflation. This is despite the fact that the 'real' outcome of the two scenarios is identical, i.e. a 2% reduction in real spending power. Thus it seems that holding real change constant; people's reactions will be determined by nominal changes. Incredibly in some situations, a nominal increase may even offset a downward real change. Results from the studies carried out by these authors suggest that the preferences of many people are heavily affected by nominal values: 'The answers of many people indicate that not only do they believe themselves prone to money illusion but also that they expect others to be

affected by money illusion' (Shafir et al., 1997: 370).

'Anchoring' is a special case of 'framing' which is especially pertinent to the housing market. Holding real (replacement) cost constant, it has been shown that in times of changing relative prices, people's reactions will be determined by the change between an items current price and its historical, nominal anchor. Loss aversion thus occurs relative to a reference point, and that reference point can indeed be nominal, yielding further manifestation of money illusion. Genesove and Mayer (2001) document that investors are particularly reluctant to realise nominal losses, even if they are gaining in real terms.

It is alleged that money illusion is manifest in economic behaviour in three main ways. It is often given as a leading explanation for the phenomenon of 'sticky prices'. Money illusion can perhaps help explain why nominal values are slow to change even when inflation has caused real prices and costs to increase. A second anomaly that theorists have sought to explain is the the fact that contracts and laws are not indexed to inflation as frequently as one would expect based on the predictions of monetary theory. Indexed contracts are often only introduced very slowly as inflation picks up, and even more startlingly, partially disappear when inflation slows down. Governments frequently use contracts that are not indexed, or only partially so. Courts do not actually treat inflation the same as any other unexpected event, which destroys the value of a contract (Lejonhufvud, 1977). Thirdly, money illusion is evident in social discourse and the media. Even in familiar contexts and among people who, on one level, know better – frequent newspaper articles, news stories and other sources give accounts of unadjusted costs, charitable donations and salaries across time (Fehr and Tyran, 2007).

Why Have Many Economists Rejected the Idea?

Recognition of the possibility of money illusion has a long standing in economics. Indeed in 1928, Irving Fisher dedicated an entire book to it. That is not to say that the theory has always been held in high esteem among monetary economists. In actual fact the converse is true. Large circles of monetary economists, who maintain that agents act rationally, heavily dispute the existence of money illusion. Subscribing to nominal values in deference to real ones is clearly not rational, but may well happen irrespectively.

Nevertheless, developing equilibrium models that account for money illusion goes against the grain of 'rational' modelling. Commenting on the prevailing attitudes amongst professional economists, Tobin states that: 'An economist can, of course, commit no greater sin than to assume money illusion'

(Tobin, 1972: 5). The same economist said that a great way to ensure that an article would not be published in one of the 'prestigious' economic journals, was to associate the article or indeed its author with money illusion.

Money illusion has been anathema to the profession for a number of decades. The index of *The Handbook of Monetary Economics* by Ben Friedman and Frank Hahn (1990), for example, does not even mention the term. The intuition behind rational economic agents rejecting money illusion is two-fold. Firstly, the objective function of the individual must depend only on real magnitudes. Secondly, people must perceive that purely nominal changes do not affect their opportunity set: i.e. people have to understand that equi-proprtionate changes in all nominal magnitudes leave their real constraints unchanged. Whether or not people are, in fact, able to pierce the veil of money, i.e. whether they understand that purely nominal changes leave their objective circumstances unchanged is at the heart of the money illusion question.

The ambivalence with which the profession regards the idea of money illusion is probably best represented in Howitt's entry in *The New Palgrave Dictionary on Economics*:

> 'The absence of money illusion is the main assumption underlying the long run neutrality of money proposition of the quantity theory of money ….. Many economists have reacted adversely to explanations based on such illusions, mainly because illusions contradict the maximising paradigm of microeconomic theory and partly because invoking money illusion is often too simplistic an explanation of phenomena that do not fit well into the standard equilibrium mould of economics… the assumption is frequently invoked and frequently resisted…' (Howitt, 1987: 518).

How Does Money Illusion Confuse 'Rational' Economic Agents?

So have supposedly rational economic agents been duped by relying excessively on nominal valuations? Explanations of money illusion generally describe it in terms of heuristics. Nominal prices provide a convenient 'rule of thumb' for determining value, and real prices are only calculated if they seem highly salient. Such cases might include the signing of very long contracts, or in a period of very high inflation. A number of authors point to the 'ease, universality and salience' of the nominal representation, and the sophistication of the decision maker (Blinder, 2000; Patinkin, 1969; Shafir et al., 1997).

People attend to nominal value because they are salient, easy to use, and

in many cases provide a reasonable estimate of real worth. Furthermore, it fits with our generl perception that most objects around us, particularly units of measurement, do not regularly change. Fundamentally, it is considered easier and more natural to think in nominal rather than real terms. This tendency is likely to continue despite economists' attempts to educate the public (Fisher, 1928). The persistence of money illusion indicates that learning fails to eliminate this monetary phenomenon. Salience is not as important as accuracy however, and using nominal values in favour of real ones can clearly be highly misleading.

Empirical Research on Money Illusion

Irving Fisher was most likely the first person to use the term 'money illusion', in his 1928 book of the same name. To gather information for his research, Fisher took the logical step and conducted interviews in post-war Germany, a country suffering from huge price-level problems. He found significant evidence that people were suffering from money illusion (Fisher, 1928). Once again using Germany as their subject, Stefan Boes and Markus Lipp used a test is based on people's self-reported satisfaction with their income. In the absence of money illusion income satisfaction should remain unchanged if commodity prices and nominal income increase or decrease in the same proportion (Boes and Lipp, 2006). If, on the other hand, a proportional increase in prices and nominal income increases subjective wellbeing, then we have evidence for money illusion. Their findings overwhelmingly suggest the widespread existence of money illusion.

Shafir et al. (1997) have provided compelling evidence for the existence of the phenomenon, showing its affect on behaviour in a variety of experimental and real world situations. Money illusion would be observed if, in the presence of inflation, nominal accounting affected real decisions, a possibility recognised by Fischer and Modigliani (1978). Moreover, with changing relative prices, the effect of past nominal values on purchase or sales decisions would be tantamount to money illusion, even in the absence of inflation. This could manifest itself in a reluctance to sell a house, shares or other assets, which result in a nominal loss, and also in a reluctance to accept nominal wage cuts.

Money Illusion and the Stock Market

Several studies suggest a negative correlation between nominal stock returns and inflation (Litner, [1975]; Fama and Schwert, 1977). This appears puzzling since the Fisher relation implies that nominal rates should move one for one with expected inflation. One possible interpretation is that since inflation proxies for future economic conditions, higher inflation is associated with a bleak economic outlook (Fama and Schwert, 1977). Modigliani and Cohn (1979) used a different approach, basing this negative correlation on money illusion. They claimed that prices significantly depart from fundamentals since investors make two 'inflation-induced judgement errors'. Firstly, they tend to capitalise equity earnings at the nominal rate rather than at the real rate of interest. Secondly, they fail to realise that a firm's corporate liabilities depreciate in real terms. Hence stock prices are too low during inflationary periods. This idea has become known as the 'Modigliani-Cohn' hypothesis and has become the basis for further studies on the effects of money illusion on the stock market (see Cohen et al., 2005).

Money Illusion and the Housing Market

Money illusion can also have profound effects on the housing market, as discussed by Brunnermeier and Julliard (2006). The authors point out that a reduction in inflation can fuel a boom in house prices. For example, investors who formulate their decision on whether to buy or rent a house by comparing rent and mortgage payments are not taking into account the fact that inflation reduces the real cost of future repayments: 'they mistakenly assume that nominal and real interest rates move in lockstep' (Brunnermeier and Julliard, 2006: 2). People incorrectly attribute a decrease in inflation to a reduction in the real interest rate and thus underestimate the future costs of mortgage repayments. This mistake helps to encourage people to purchase property, consequently putting upward pressure on the housing market in times of reducing inflation.

By trying to isolate the fundamental components of house-price changes, such as land prices, economic growth, and property taxes, the authors aim to distinguish between 'fundamental' factors and those which are influenced purely by inflation. The close link between inflation and large run-ups in housing prices may well be attributable to money illusion. As stated above, inflation may lead people to erroneously believe that real interest rates on borrowings are lower than they actually are – thus confusing nominal and real terms.

So the current depression in the UK and Irish housing markets (where

sub-prime loans were not nearly as common as in the US) may indeed be partly attributable to the effects of money illusion. Recent increases in Irish inflation rates are now hurting homeowners and potential buyers as nominal interest rates are finally beginning to catch up with real ones.

Money Illusion in the Labour Market

In the short run, it is possible that workers may be more responsive to money wages than to real wages. Depending on the speed with which reliable information on wages and the price level becomes available, people may respond more quickly to changes in nominal wages. Only in the long run is it likely that the real wage is the most influential factor in the labour markets.

Money illusion is perhaps the most undeniable explanation of the 'wage-rigidity' phenomenon. Even in countries with persistently high levels of unemployment, nominal wages tend to be 'sticky' and unconducive to downward shifts. According to J.M. Keynes, the best way to expand employment is to reduce real wages by pushing up the price level. Economic agents, prone to money illusion, will fail to take account of their newly reduced spending power. If workers use nominal wages as a reference point when evaluating wage offers, firms can keep real wages relatively lower in a period of high inflation as workers accept high nominal wage increases. These lower real wages would allow firms to hire more workers in periods of higher inflation.

If workers are not prepared to accept a reduction in their real wages, brought about by wage cuts, then why should they accept a similar reduction engineered through price increases? According to Keynes, economic agents are prone to irrationality, so he actually advocated the controlled increase of inflation. In other words, not only did Keynes believe in the existence of money illusion, he actually supporteed its perpetuation as a method of helping the government finance the war effort! (Keynes, 1940)

Further analysis of the Phillip's curve, long regarded as the empirical embodiment of the Keynesian theory of inflation, occupies a more dubious position in latter macroeconomics. Nevertheless, money illusion may well be involved in the formation of inflation expectations. An eminent Neo-Keynesian, wrote:

> 'The Phillips curve idea is in a sense a reincarnation in dynamic guise of the original Keynesian idea of money illusion in the supply of labour. The Phillips curve says that increased money wages are in some significant degree prized in themselves even if

> they do not result in equivalent gains in real incomes' (Tobin, 1967 cited in Fehr and Tyran, 2001: 1244).

It has been suggested that, as a result of money illusion, the negative relation between inflation and unemployment as described by the Phillips curve may indeed hold – contrary to modern revisions of the model (Fehr and Tyran, 200).

Conclusion

> 'A small amount of individual level irrationality can have large aggregate effects'(Akerlof and Yellen, 1985: 139).

It would be wrong to suggest that everybody is guilty of placing excessive importance on nominal values. Some economists contend that people do understand the importance of basing decisions solely on real values. But, as the above quote from Akerlof and Yellen implies, even small illusory effects on an individual level can have large aggregate effects.

Perhaps the greatest problem facing the 'rational' economists battling the concept of money illusion, is the promulgation of nominal values throughout the media and everyday social discourse. As stated at the outset, nominal values have become the bedrock of 'popular', 'lay-man's economics.

The large number of empirical studies in this area, some of which have been discussed in this essay, offer indisputable evidence for the existence and indeed prevalence of money illusion in today's society. The phenomenon is not a purely economic one, having also been extensively studied in the realm of psychology and behavioural finance. The great John Maynard Keynes not only acknowledged money illusion but actually advocated using the phenomenon as a method of financing the war! It is hard to disagree with Blinder when he states that:

> 'In fact, I am persuadable – indeed, pretty much persuaded – that money illusion is a fact of life'(Blinder, 2000: 54).

Bibliography

Akerlof, G.A. and Yellen, J.L. 1987. 'Rational Models of Irrational Behaviour'. *The American Economic Review* 77:2:137-142.

Blinder, A. 2000. 'Comment on Near-Rational Wage and Price Setting and the Long-Run Phillips Curve'. *Brookings Papers on Economic Activity* 2000:2:50-55.

Boes, S., Lipp, M. and Winklemann, R. 2006. 'Money Illusion Under Test'. *Economics Letters* 94:332-337.

Brunnermeier, M.K. and Julliard, C. 2006. *Money Illusion and Housing Frenzies*. London: Centre for Economic Policy Research.

Cohen, R.B., Polk, C., and Vuolteenaho, T. 2005. 'Money Illusion in the Stock Market: The Modigliani-Cohn Hypothesis'. *Quarterly Journal of Economics* 120:2:639-68.

Fama, E.F. and Schwert, G.W. 1977. 'Asset Returns and Inflation'. *Journal of Financial Economics* 5:2:115-46.

Fehr, E. and Tyran, J.R. 2001. 'Does Money Illusion Matter?' *The American Economic Review* 91:5:1239-62.

Fehr, E. and Tyran, J.R. 2007. 'Money Illusion and Coordination Failure'. *Games and Economic Behaviour* 58:2:246-268.

Fischer, S. and Modigliani, F. 1978. 'Towards an Understanding of the Real Effects and Costs of Inflation'. *Weltwirtschaftliches Archive* 114:810-833.

Fisher, I. 1913. 'A Compensated Dollar'. *The American Economic Review* 27:2:213-235.

Fisher, I. 1928. *The Money Illusion*. New York: Adelphi.

Genesove, D. and Mayer, C. 2001. 'Loss Aversion and Seller Behaviour: evidence from the housing market'. *Quarterly Journal of Economics* 116:4:1233-1260.

Howitt, P. 1987. 'Money Illusion', in Eatwell et al. *The New Palgrave Dictionary of Economics*. New York:W.W. Norton.

Keynes, J.M. 1940. 'How to Pay for the War: A radical plan for the Chancellor of the Exchequer'. *Round Table War Pamphlets* No.1.

Leijonhufvud, A. 1977. 'Costs and consequences of inflation,' in G.C. Harcourt (ed.) *The Microeconomic Foundations of Macroeconomics*. London: Macmillan.

Modigliani, F. and Cohn, R.A. 1979. 'Inflation, Rational Valuation and the Market'. *Financial Analysts Journal*. 38: 24-44.
Patinkin, D. 1965. *Money, Interest and Prices*. New York: Harper and Row.

Shafir, E., Diamond, P. and Tversky, A. 1997. 'Money Illusion'. *Quarterly Journal of Economics* 112:2:341-74.

Tobin, J. 1972. 'Inflation and Unemployment'. *The American Economic Review* 62:1:1-18.

GERMAN INFLATION AND THE MONEY SUPPLY, 1919-1923

DERVLA BRENNAN

Senior Sophister

Hyperinflation haunts the dreams of many a central banker. The German experience post World War I is perhaps the most frequently cited example of this inflationary nightmare occurring in reality. In this paper, Dervla Brennan provides some invaluable advice for any burgeoning economist seeking to become the 'next' Alan Greenspan. A number of theories of hyperinflation are extensively discussed. As ever, the difficulty is in separating economic facts from politics. The growth of the money supply is considered to have been the driving force behind the inflationary crisis, but whether the supply was endogenous or exogenous is a major point of contention.

Introduction

> 'A government can live for a long time, even the German government or the Russian government, by printing paper money…A government can live by this means when it can live by no other'.[1]

According to Milton Friedman, 'Inflation is always and everywhere a monetary phenomenon'. Inflations and hyperinflations, by definition and historically, are periods of rapid trend increases in price combined with a decreasing demand for real money balances (Holtfrerich, 1986). In the 1920s the word inflation came to signify money growth rather than a rise in prices (Webb, 1985). Indeed the German inflation from 1919 to 1923 saw an immense increase in the circulation of paper money. By the end of November 1923 there were 400,267,640,302 billion notes in circulation compared with 663,200 billion in August 1923 (Sargent, 1981). This was accompanied by a rapid depreciation of the mark. In

[1] J.M. Keynes. A Tract on Monetary Reform (1923)

July 1919, one US dollar could be bought at 14 marks. By November 1923, the exchange rate was 4,420,000,000,000 marks to the dollar (Berghahn, 1987). It was claimed by the Allies that the Reichsbank was purposely expanding the money supply and thus depreciating the mark in order to indicate their inability to pay reparations. Meanwhile it was claimed in Germany that the currency depreciation, arising from disequilibrium of the balance of payments, had caused the expansion of the money supply and the inflation.

What was the primary determinant of Reichsbank money supply policy? This paper will begin with an examination of the Reich's fiscal policy during and after the war. A summary will then be presented of two competing theories that seek to explain the inflation that existed in Germany at the time: the 'Balance of Payments' theory and the 'Quantity' theory. The crux of this paper concerns Cagan's (1956) seminal work on the demand for money during the hyperinflation in Germany. Cagan's paper reopened the debate on what determined the money supply process in Germany post World War I. Ultimately this leads on to an examination of whether the money supply was exogenous as determined by the Reichsbank, or endogenous as determined by the actions of other agents in the economy, with the Reichsbank playing a passive role.

Reich Fiscal Policy, 1914-1923

Germany relied on printing money instead of levying taxation to finance their participation in World War I. This is the most inflationary form of war finance (Holtfrerich, 1986:107). However, inflation is a form of tax which acts unbeknownst to the public. As Keynes remarks: '[inflation] is the form of taxation which the public find hardest to evade and even the weakest government can enforce, when it can enforce nothing else' (Keynes, 1923: 37). The economic cost of an inflationary tax is the weakening power of the currency to carry out its functions (Holtfrerich, 1986: 121).

Following the war the main economic question facing the German government was whether they should restore the mark to its pre-war value or maintain the value it had obtained by the end of the war (ibid). Germany chose the latter, continuing with a policy of price inflation and exchange rate depreciation, and thus also the inflation tax. This particular policy option is reflective of the centrality the budget deficit had come to hold in currency debates after the war (ibid). Ideally a tax reform was needed in order to meet the Reich's expenses. However, the government was too weak to ask such sacrifices from its people, especially after having lost the war. Increasing the money supply was the

easiest means of meeting fiscal requirements (Bresciani-Turroni, 1937). Haller remarks: 'If the state did cause the inflation, it did so in self-defence' (Holtfrerich, 1986: 137).

Furthermore, Germany's total financial obligation under the Treaty of Versailles was for a long time unknown, leaving fiscal policy uncertain, and this uncertainty destabilized the value of the mark (Sargent, 1981). Threatened by allied penalties on the one hand and by collapse of the Reich on the other, no German government could have fulfilled the reparation demands in any way other than by printing money (Holtfrerich, 1986). Germans have always cited the terms of the Treaty of Versailles as the reason for the inflation. Indeed both 'Balance of Payments' and 'Quantity' theorists saw reparations as the disequilibrating factor, acting either on the balance of payments or on the fiscal balance, and thereby hastening the inflation and currency depreciation (ibid). This prevalent belief prevented any consideration of monetary reform, which may have impeded progress in tackling the inflation.

In November 1923 prices suddenly stopped rising, the mark suddenly stopped depreciating, the government put an end to borrowing from the central bank, the budget swung into balance and consequently the inflation came to an end. Sargent argues that a permanent change in the fiscal and monetary policy regimes was what was required to end the inflation (Sargent, 1981). Indeed the abrupt change outlined above was facilitated by two such simultaneous measures. The monetary reform involved the creation of a new currency; the retenmark,[2] and a new note issuing bank; the Retenbank. This bank was restricted both in the total volume of retenmarks it could circulate and, more importantly, on the amount of credit it could extend to the government for financing the deficit. The fiscal reform entailed implementing a new tax system and curtailing expenditure. Under the Dawes plan Germany's reparations payments were reviewed and reduced to a more manageable sum (Sargent, 1981). Thus earlier attempts to end the hyperinflation had failed because the government did not revise their fiscal policy of financing the deficit through increments of an unbacked money supply.

Balance of Payments Theory versus Quantity Theory

During the war itself, and especially in the aftermath, a debate broke out in political and academic circles regarding two theories of inflation. 'Balance of

[2] 1 retenmark = 1 trillion paper marks.

Payments' theorists cited the passivity of the German external balance caused by reparations and the other exactions under the Treaty of Versailles as the cause of the depreciation of the mark and the ensuing inflation. The depreciation of the mark was thus not the consequence but the cause of the budget deficit, money supply expansion and the inflation (Bresciani-Turroni, 1937; Holtfrerich, 1986). According to Helfferich:

> 'First came the depreciation of the German currency by the overburdening of Germany with international liabilities and by the French policy of violence…inflation is not the cause of the rise in prices and of the depreciated currency, but the latter is the cause of the higher prices and of the greater volume in the issue of paper money…'(Laidler and Stadler, 1998: 820).

'Quantity' theorists, on the other hand, saw the continued issues of paper money to finance the budget deficit as the fundamental cause of the inflation and currency depreciation (Bresciani-Turroni, 1937; Holtfrerich, 1986). Robinson argues that theoretical discussion of the German inflation was for some time clouded by political prejudices, with the German writers blaming reparations and the collapse of the exchange rate, and the Allies blaming the budget deficit and creation of money (Robinson, 1938). It is claimed that while the Reichsbank publicly held the 'Balance of Payments' theory as the root cause of the inflation, they essentially adhered to the 'Quantity' theory in their confidential correspondence with the government.

Recent Theories of Hyperinflation

Cagan's 1956 Model

> 'Hyperinflation provides a unique opportunity to study monetary phenomena' (Cagan, 1956: 25).

In Germany, hyperinflation is defined as beginning in August 1922 and ending in November 1923[3]. Real cash balances fell over the whole period of hyperinflation but they fluctuated drastically from month to month. Cagan's paper proposes and tests a theory that accounts for this erratic behaviour of real cash

[3] Cagan defines hyperinflation as: 'beginning in the month the rise in prices exceeds fifty per cent and as ending in the month before the monthly rise in prices drops below that amount and stays below for at least a year'(Cagan,1956: 25).

balances (ibid). By allowing for lags[4] and assuming an exogenous money supply, Cagan hypothesizes that variations in the expected rate of inflation[5] account for variations in real cash balances during hyperinflation. This is the case only where expectations are formed adaptively: that is by extrapolating past rates of inflation into the future. This implies a dynamic process in which the hyperinflation of prices was caused by both past and current changes in the quantity of money. Cagan concludes that: 'domestic monetary factors alone explain hyperinflation' (ibid: 90). He attributes this tremendous increase in money and prices to the fiscal needs of the government:

> 'Issuing money was a method of raising revenue by a special kind of tax – a tax on cash balances…All that is required is to spend newly printed notes. The resulting inflation automatically imposes a tax on cash balances by depreciating the value of money' (ibid: 78).

However, Cagan's model is incapable of explaining the final months of the German hyperinflation, for which it predicts lower real money balances than those actually observed from August through November 1923.[6] He offers two explanations that may account for these outlying observations. First, he hypothesizes that rumours of currency reform led people to expect that prices would stop rising after a certain number of months. This encouraged people to hold higher real cash balances than they would otherwise have held in view of the current expected rise in the price level. The second explanation Cagan offers is that he had used the wrong functional form in deriving his money demand function[7] (ibid). Cagan's seminal paper on the demand for money under hyperinflation has thus triggered an investigation into the German hyperinflation. In particular, various attempts have been made to explain the final months of 1923 when, contrary to conventional theory, real money balances increased despite accelerating inflation (Flood and Garber, 1980).

---

[4] 'The large changes in the balances during hyperinflation correspond to large changes in the rate of price change with some delay, not simultaneously.' (Cagan, 1956: 88)

[5] This is equivalent to the rate of depreciation in the value of money or a decline in the purchasing power of nominal money balances.

[6] The period of maximum inflation and maximum money creation.

[7] Frenkel (1977) in his study of the German hyperinflation rejects the possibility that Cagan had used the wrong functional form in deriving money demand.

Flood and Garber: Hyperinflation and Rational Expectations

Flood and Garber (1980) combine Cagan's model of hyperinflation with rational expectations instead of adaptive expectations in order to explain the final months of the German hyperinflation. They extend the standard monetary theory and include observations from August through November 1923. The purpose of this was to test Cagan's hypothesis that agents expected a monetary reform toward the end of the hyperinflation and thus began to hold higher cash balances despite increasing inflation. True to the spirit of rational expectations, Flood and Garber argue that when people anticipated a permanent change in the money supply regime, they changed their expectations of future inflation to reflect the change in monetary policy. They calculated the probability of monetary reform with the following result: '…our measure of the probability of monetary reform…hit its peak during the very week [November 15, 1923] commonly asserted to mark the beginning of the reform' (Flood and Garber, 1980: 48). Thus they conclude that expectations of monetary reform were the factor responsible for increasing real cash balances.

Webb: Endogenous Money Supply

Webb (1985, 1986) asserts: 'the primal cause of the money growth and the whole inflation was the growth of the government debt' (Webb, 1985: 490). In his analysis, Webb treats the money supply as endogenous rather than exogenous. He argues that one should focus on the determinants of money supply rather than money demand. The money supply depended on the government debt and how much of this debt the public decided to monetize. This fraction of the debt to be monetized was in turn influenced by inflationary expectations and on public confidence in German finances. Thus government debt and inflationary expectations indicating an endogenous money supply explain the money supply better than the assumption of exogeneity (Webb, 1986). Webb contends that it was positive 'fiscal news' combined with rumors of monetary reform in August 1923 that increased the possibility that the government would cease to run deficits. A new government came to power mid-August and succeeded in increasing revenue despite the increasing inflation. In September, the government declared an end to passive resistance which further lowered government expenditures (ibid). In addition, from August 1923 it seemed likely that the Reichsbank was planning to institute a new policy regime. The Reichsbank's refusal to discount commercial bills of more than a month's duration was combined with an obligation to repay in terms of gold. They also sent a private memorandum to the government informing them that after the end of the year, the Reichsbank would cease to monetize any more government debt. This was the first time the national bank threatened to use its supposed autonomy

from the government. While this announcement was not made public, Webb contends that it must have been known in the upper circles of the business community that sat on the Direktorium and Aufsichtsrat of the Reichsbank (ibid). Thus both signals of fiscal reform and monetary reform played a role in lowering inflationary expectations towards the end of the hyperinflation.

Tullio's Dynamic Model

The debate at this point rests on the question of whether the money supply during the inflation was exogenous or endogenous. Tullio (1995) develops a dynamic model of the German hyperinflation, simultaneously explaining prices, the exchange rate, and money supply. In his model the money supply is partially endogenized, and expressed as a function of nominal income and of the deviation of actual from potential output. The stability of the model indicates that the German inflationary process was primarily caused by fiscal deficit and excessive monetary growth (Tullio, 1995). Tullio argues that there is some evidence indicating that the Reichsbank passively financed increases in prices at times of low economic activity in order to avoid a recession, allowing for the partial endogeneity of the money supply (ibid). This partial endogeneity implies the cause of the inflation was at times from prices to money. These results support the hypothesis that the Reichsbank followed this real-bills doctrine, passively financing increased demands for money and credit. In 1991, Cagan himself returned to the topic and concluded:

> 'the money stock cannot be treated as exogenous. A plausible way to endogenize the money stock is to model the revenue needs of the government for an inflation tax' (Laidler and Stadler, 1998: 822).

Conclusion

The growth of the money supply is a prominent feature in every account of the German inflation. While money growth was frequently cited as a consequence and sometimes as the cause of inflation, recent models of inflation, and in particular hyperinflation, treat the growth of money as an immediate and often as an exogenous cause of inflation. In these models inflationary expectations have come to play an important intermediary role (Webb, 1985). Inherent in these formulations is a denial of the possibility that inflation has a momentum of its own. Rather it is the long-term government policy of persistent budget deficits and high rates of money creation which gives momentum to the inflation rate (Sargent, 1981). However, these models that accord a role to expectations in the

inflationary process are by no means perfect. There is, as yet, no ideal way to model expectations. While the rational expectations formulation is certainly useful, it is based on very strong assumptions and so comes with certain limitations.

Kiguel (1989) notes that there was an element of bidirectional causation during the inflation. High rates of inflation decreased the real value of government revenues, thus increasing their financial requirements. The Government then financed this deficit in revenue by increasing the money supply, reinforcing the inflationary process (Kiguel, 1989). Thus at times the inflation did run from prices to money. What is certain is that the inflation was caused by the acceleration of money growth to finance the large accumulation of government debt. This relentless printing of money was the only means of survival for a weak republican government burdened with the charge of having accepted the humiliating Treaty of Versailles. The exact nature of the money supply still remains a topic of debate. However, contentions of an endogenous money supply are dubious in the German case. The Weimar government made a conscious decision to finance the budget deficit by printing money, thus implying an exogenous money supply.

Bibliography

Berghahn, V. R. 1987 [1982]. *Modern Germany: Society, Economy and Politics in the Twentieth Century*. Cambridge: Cambridge University Press.

Bresciani-Turroni, C. 2003 [1937]. *The Economics of Inflation: a Study of Currency Depreciation in post-war Germany.* London: Routledge.

Cagan, P. 1956. 'The Monetary Dynamics of Hyperinflation' in M. Friedman (ed.) *Studies in the Quantity Theory of Money*. Chicago: University of Chicago Press.

Flood, R.P. and Garber, P.M. 1980. 'An Economic Theory of Monetary Reform'. *The Journal of Political Economy* 88:1:24-58.

Frenkel, J. A. 1977. 'The Forward Exchange Rate, Expectations, and the Demand for Money: The German Hyperinflation'. *The American Economic Review* 67:4:653-670.

Holtfrerich, C. 1986. The German Inflation 1914-1923: *Causes and Effects in International Perspective*. Berlin: de Gruyter.

Keynes, J. M. 1971 [1923]. *The Collected Writings of John Maynard Keynes Vol.4: A Tract on Monetary Reform*. London: Macmillan.

Kiguel, M. A. 1989. 'Budget Deficits, Stability, and the Monetary Dynamics of Hyperinflation'. *Journal of Money, Credit and Banking* 21:2:148-157.

Laidler, D.E.W. and Stadler, G.W. 1998. 'Monetary Explanations of the Weimar Republic's Hyperinflation: Some Neglected Contributions in Contemporary German Literature'. *Journal of Money, Credit and Banking* 30:4:816-831.

Robinson, J. 1938. 'Review of Bresciani-Turroni's The Economics of Inflation'. *Economic Journal* 48:191:507-513.

Sargent, T. J. 1981. 'The Ends of Four Big Inflations' in R.E. Hall (ed.) *Inflation: Causes and Effects*. Chicago; London: University of Chicago Press.

Tullio, G. 1995. 'Inflation and Currency Depreciation in Germany, 1920-1923: A Dynamic Model of Prices and the Exchange Rate'. *Journal of Money, Credit and Banking* 27:2:350-362.

Webb, S. 1985. 'Government Debt and Inflationary Expectations as Determinants of the Money Supply in Germany 1919-1923'. *Journal of Money, Credit and Banking* 17:4:479-492.

Webb, S. 1986. 'Fiscal News and Inflationary Expectations in Germany After World War I'. *Journal of Economic History* 46:3:769-794.

CHALLENGING MERCANTILISM: THE IMPACT OF DAVID HUME ON THE EVOLUTION OF MONETARY THOUGHT

ALEXANDER TOFT

Senior Sophister

Despite its somewhat irrational and outdated nature, mercantilist sentiment is very much alive today. In times of economic slowdown, it often bubbles to the surface of popular economic discourse, threatening to make an unwelcome return to politics and policy. The spirits of mercantilism and protectionism have always gone very much hand in hand. In this paper, Alexander Toft provides an excellent appraisal of David Hume's monetary thought, in particular examining his attack on the 17th century mercantilist monopoly over economic discourse.

Introduction

The history of economic thought can be viewed as a series of ideological battles, and key amongst them must rank the debate surrounding the notion that the government has a role to play in ensuring a more favorable balance of trade – what is commonly called mercantilism. Adam Smith is often credited with initiating this debate when he attacked the concept in *The Wealth of Nations* (Smith, 1776). However, this paper will examine the work of another original detractor from mercantilist thinking, David Hume. A philosopher and historian first and foremost, Hume was also an eminent economist. This paper will argue that through his analysis of the influence of money on inflation and the balance of payments, Hume played a pioneering role in challenging the mercantilist monopoly on economic thought. He contributed towards the development of an alternative theory which centered on free trade, as opposed to a 'fear of goods' and 'love of money'. To this end, the following paper will outline and analyze Hume's work on monetary policy, breaking it into three main components: his view of money, the interest rate and the balance of payments. Finally, the impact of these ideas on the evolution of economic thought will be considered.

Mercantilism

Before we begin, it is important to define what exactly mercantilism is. According to Speigel, the concept has its roots in the 'bullionist' regulations of medieval times, but took hold as a more formal doctrine when developed and articulated by writers such as Sir Thomas Mun in the early seventeenth century (Speigel 1991). It is undoubtedly a rather general concept with many possible interpretations. However, perhaps its defining feature is the belief that monetary balances, as the key source of a nation's wealth, are crucial to the state and should thus be supported by protectionist policies. More specifically, it was (and still is) a view which held:

> 'that a favorable balance of trade was of transcendent importance for a nation's political economy; that a low rate of interest was caused by the bullion surplus earned by a favorable trade balance; and that both at home and abroad the control over the use of resources should be regarded as a zero-sum game in which what one country gained another lost' (Rostow, 1990: 21).

This attitude is what Hume was contending against. Keeping this in mind, we are now ready to discuss his work.

Hume's Monetary Thought

Hume's economic writings constitute only a small fraction of his work, but nonetheless cover a wide range of topics, from commerce to public credit. These writings can be largely derived from the essays that make up his *Political Discourses* (1752). In this section we will limit our analysis to his work on money, interest rates, and the balance of payments, examining each in turn.

Of Money

In the essay *Of Money*, Hume challenges traditional mercantilist thinking by claiming that 'money is not, properly speaking, one of the subjects of commerce; but only the instrument which men have agreed upon to facilitate the exchange of one commodity for another' (Hume, 1752 cited in Rotwein, 1998: 33). Its value is therefore 'fictitious' and nothing more than 'the representation of labor and commodities', which are the true sources of economic power (ibid: 37). A key implication of this definition is that money is neutral; changing its quantity affects nothing but the price level (Niehans, 1990: 52). Consequently, in a closed

economy, 'it is evident, that the greater or less plenty of money, is of no consequence' (Hume, 1752 cited in Rotwein, 1998: 33). Money does not constitute the wealth of a nation or the 'wheels of trade' as mercantilism often holds, but is simply 'the oil which renders the motion of the wheels more smooth and easy' (ibid: 33).

Hume does not suggest, however, that money is irrelevant. In an open economy, the inflationary effects of money give it significance. The logic is as follows: when a country grows rich through trade and specie accumulation it experiences higher prices, including higher wages. The resulting cost pressures lead manufacturers to leave the country and relocate elsewhere, 'whither they are allured by the cheapness of provisions and labour' (ibid: 34, 35). Thus:

> 'the dearness of every thing, from plenty of money, is a disadvantage, which attends an established commerce, and sets bounds to it in every country, by enabling the poorer states to undersell the richer in all foreign markets' (ibid: 34,35).

Therefore, money can have the effect of undermining a country's ability to engage in profitable foreign trade.

Hume provided an important break from mercantilist thought by suggesting that the value of money is ultimately fictitious and that an increase in the money supply has inflationary effects, which may be damaging to a country's economy. However he did acknowledge that:

> 'in every kingdom, into which money begins to flow in greater abundance than formerly, everything takes a new face; labor and industry gain life; the merchant becomes more enterprising, the manufacturer more diligent and skilful, and even the farmer follows his plough with greater alacrity and attention'(ibid: 37).

What Hume identifies here is the seemingly paradoxical short-term non-neutrality of money, despite its long-term neutrality. The explanation for this phenomenon can be considered one of his greatest contributions to economics – that an increase in the money supply is not immediately followed by an increase in the price level. Rather there exists an interval between an increase in the money supply and a rise in prices during which money gradually disperses from the hands of a few merchants, stimulating industry and increasing both output and employment. Nevertheless, in the long-run, 'the whole effect [exhausts] itself out in neutral price increases' (Niehans, 1990: 54). Such a theory leads Hume to

advocate a monetary or trade policy which maintains a gradually increasing money supply, arguing that:

> 'the good policy of the magistrate consists only in keeping [money], if possible, still encreasing; because, by that means, he keeps alive a spirit of industry in the nation, and encreases the stock of labour, in which consists all real power and riches' (Hume, 1752 cited in Rotwein, 1998: 39, 40).

Of Interest

Hume's thoughts on the interest rate derive from his views on money. In particular, his belief that money possesses a fictitious value allows him to dismiss the notion, later defended by J.M. Keynes in *The General Theory* (1937), that a low interest rate is a result of a high money supply. Indeed, in *Of Interest*, Hume suggests that the rate of interest is 'not derived from the quantity of precious metals' at all (ibid: 48).

However, he does not dismiss the actuality that 'nothing is esteemed a more certain sign of the flourishing condition of any nation than the lowness of interest' (ibid: 47). Rather it is argued that low interest rates have sources other than money. These sources include a low demand for borrowing, an abundance of riches to supply that demand, and small profits arising from commerce – all of which derive from 'the encrease of industry and commerce, not of gold and silver' (ibid: 49). In this way, interest rates reflect the supply and demand of real capital, which in large part are dependent upon 'the habits and way of living of the people' (Spiegel, 1991: 211).

In proposing such a theory, Hume provided a unique counterpoint to mercantilist thinking. Economists such as Sir William Petty had, in some ways, already managed to play down the role of money as a driving force in the economy. However, Hume more specifically demonstrated that money is unable to produce low interest rates, undermining the logic that Keynes would eventually use to defend mercantilist policies (Higgs, 1926: 342).

Of the Balance of Trade

For the purposes of this paper, *Of the Balance of Trade* represents the culmination of Hume's economic writings as it contains his most direct attack on mercantilism. In it, he dismisses as 'groundless apprehension' what he perceives to be mercantilist 'jealousy with regard to the balance of trade, and a fear, that all their gold and silver may be leaving them' (Hume, 1752 cited in Rotwein, 1998:

61). This reasoning can be traced back his theories on money and is explained in the following key passage outlining the logic behind the price-specie flow mechanism:

> 'Suppose four-fifths of all the money in Great Britain to be annihilated in one night, and the nation reduced to the same condition, with regard to specie, as in the reigns of the Harrys and the Edwards, what would be the consequence? Must not the price of all labour and commodities sink in proportion, and everything be sold as cheap as they were in those ages? What nation could then dispute with us in any foreign market, or pretend to navigate or to sell manufactures at the same price, which to us would afford sufficient profit? In how little time, therefore, must this bring back the money which we had lost, and raise us to the level of all neighbouring nations? Where, after we have arrived, we immediately lose the advantage of the cheapness of labour and commodities; and the farther flowing in of money is stopped by our fullness and repletion' (ibid: 63).

The price-specie flow mechanism constitutes perhaps Hume's most important contribution to economic thought and can be summarized as follows: in an open economy, a fall in the money supply leads to a fall in prices which boosts exports, resulting in an inflow of money that eventually raises the price level back to its original position. So what Hume describes is 'an automatic feedback control mechanism' through which the amount of specie in each nation automatically reverts towards an equilibrium where exports and imports are in balance (Niehans 1990: 55). This monetary influence on the balance of payments contradicts the mercantilist notion that a country can maintain a continuously favorable balance of trade. This implies that the amount of specie a nation can accumulate does not derive from protectionist policies, but is dependent upon a nation's level of economic development, which is determined in turn by its population and the 'spirit of industry of its people' (Eatwell et al, 1987: 693).

It is important to recognize, however, that in *Of the Balance of Trade* Hume does not simply set out to discredit the protectionist policies that may arise from a mercantilist obsession with money. Rather he is attempting to provide a robust defense of free trade (Rostow 1990: 21). In his own words:

> 'From these principles we may learn what judgment we ought to form of those numberless bars, obstructions, and imposts, which all nations of Europe, and none more then England, have put upon

> trade; from an exorbitant desire of amassing money, which never will heap up beyond its level, while it circulates; or from an ill-grounded apprehension of losing their specie, which never will sink below it. Could any thing scatter our riches, it would be such impolitic contrivances. But this general ill effect, however, results from them, that they deprive neighbouring nations of that free communication and exchange which the Author of the world has intended, by giving them soils, climates, and geniuses, so different from each other' (Hume, 1752 cited in Rotwein, 1998: 75).

According to Hume, trade is by no means a zero-sum game, but rather is critical to the promotion of a country's economic development. Speigel suggests that: '[for Hume], individuals as well as nations need not fear the prosperity of their neighbours; they can only benefit from being members of a prosperous community' (Speigel, 1991: 209).

Hume's Impact

At this point we have analyzed some key aspects of Hume's monetary thought. But what impact did such thoughts have? The influence of Hume's writings on his friend and countryman Adam Smith has been well documented. In particular, his belief in economic individualism and critical attitude towards the mercantilists was shared by Smith, even if he surprisingly failed to mention Hume's price-specie flow mechanism (Spiegel 1991).

However, Hume's legacy extends even further than his own time. There are a number of striking similarities with twentieth century economists such as Milton Friedman. In the *Quantity Theory of Money* (1987), which tellingly opens with a passage from Hume, Friedman concludes that the short-run effect of a change in the money supply is primarily on output, but that the long-run effect is on the price level. It can be argued that Friedman's conclusions quite clearly resemble Hume's own pivotal notion that the inflationary effect of an increase in money supply is a gradual process, causing money to exhibit short-term non-neutrality (Wennerlind, 2005: 224). In addition, his argument in *The Role of Monetary Policy* (Friedman, 1968) that steady, moderate growth in the money supply would be good for the economy and avoid both inflation and deflation is the same conclusion Hume reaches in *Of Money*. Friedman was perhaps best known as a proponent of free-markets. Once again, Hume's explicit and implicit attacks on protectionist policies and his promotion of free-trade at a time when mercantilist thinking still ruled, makes him an important forerunner in this

respect.

But does this really afford Hume a key place in the canon of monetary thought? After all, his views on wealth, interest and his general opposition to mercantilist thinking can be said to have been preceded by economists such as Barbon, Berkeley and Massie in the same way we have argued that his writings preceded Smith's and Friedman's (Higgs, 1926: 341). This paper argues that Hume does deserve such a place because he was able to combine the ideas of his predecessors and contemporaries into a systematic whole which presented an alternative to mercantilist thought. His writings prepared the way for a new set of doctrines specifically designed to replace the traditional teachings of Adam Smith's predecessors. Indeed that 'new set of doctrines' has proved remarkably enduring. The second half of the twentieth century can be viewed as the triumph of free-trade and globalization following the struggle between a Keynesian belief in the merits of government regulation and the free market system (Yergin and Stanislaw, 1998). Hume's role as one of the original detractors from mercantilists thinking and key architect of 'classical' economics suggests he was critical to this eventual triumph.

Conclusion

The ideological debate over the merits of mercantilism is far from over. However the purpose of this paper is not to suggest that mercantilism is either right or wrong, but to highlight the critical role of David Hume in providing an alternative to the mercantilist monopoly on seventeenth century economic thought. Hume argued that the value of money is 'fictitious'. He outlined the long-term neutrality of money, in particular its inability to affect the interest rate, by demonstrating the inflationary effects of increases in the money supply. Finally, he challenged the view that foreign trade is simply 'a strategic device to produce specie' (Spiegel, 1991: 208). He proposed a relationship between money, the price level, and the balance of payments, culminating in his price-specie flow mechanism, which provided an important counterpoint to traditional mercantilist thinking. Hume will perhaps always be best remembered for his work as a philosopher and historian, but his contribution to economic thought appears no less significant.

Bibliography

Eatwell, J., Murray, M. and Newman, P. (ed.) 1987. *The New Palgrave Dictionary of Economics*. London: MacMillan Press Ltd.

Friedman, M. 1989. 'The Role of Monetary Policy'. In *The American Economic Review*, LVII: 1-17.

Higgs, H. (ed.) 1926. *Palgrave's Dictionary of Political Economy*. London: MacMillan and Co Ltd.

Niehans, J. 1990. *A History of Economic Theory: Classic Contributions, 1720-1980*. Baltimore: The Johns Hopkins University Press.

Rostow, W. W. 1990. *Theorists of Economic Growth From David Hume to the Present*. Oxford: Oxford University Press.

Rotwein, E. 1998. *David Hume: Writings on Economics*. USA: University of Wisconsin Press.

Spiegel, H. 1991. *The Growth of Economic Thought*. USA: Duke University Press.

Wennerlind, C. 2005. 'David Hume's Monetary theory Revisited: Was He Really a Quantity Theorist and an Inflationist?' *Journal of Political Economy* 33: 139-160.

Yergin, D. and Stanislaw, J. 1998. *The Commanding Heights*. USA: Simon and Schuster.

END OF THE ROAD? – EXAMINING THE CAUSES OF THE DECLINE IN CYCLING IN DUBLIN

GAVIN DOHERTY

Senior Sophister

Gavin Doherty's 'ode to the humble bicycle' is a sad and sorry tale. It is a telling indictment of our recent economic success and the resultant affluence, that the cyclist has become an endangered species in Dublin, both literally and metaphorically. Poorly planned urban sprawl and changes in attitudes have contributed to the decline in cycling. As a result, the bicycle has become an inferior good. This paper contrasts the Irish experience with that of other European nations, suggesting that if the bicycle is to be saved, countries such as the Netherlands must serve as a guide to future policy.

End of the Road for the Bicycle in Dublin?

> *'Let a man find himself, in distinction from others, on top of two wheels with a chain - at least in poor countries like Russia - and his vanity begins to swell out like his tires. In America it takes an automobile to produce this effect.'*[1]

In a world with energy shortages, the bicycle is the most energy efficient form of transport (Gordon, 1991). In a world struggling to control emissions, the bicycle emits no pollution. Over short journeys it is often the most expedient form of transport, not being constrained by traffic congestion. However the latest census has revealed that only 1.9% of workers commute by bicycle in Ireland, less than a third of the proportion cycling to work in 1986 (CSO, 2006). This reflects the downward trend in the number of cyclists, mirrored by consistent increases in the number of motorists.

The scope of this essay will be limited to urban cycle commuting and thus will centre on Dublin; however nationwide trends will also be discussed. The

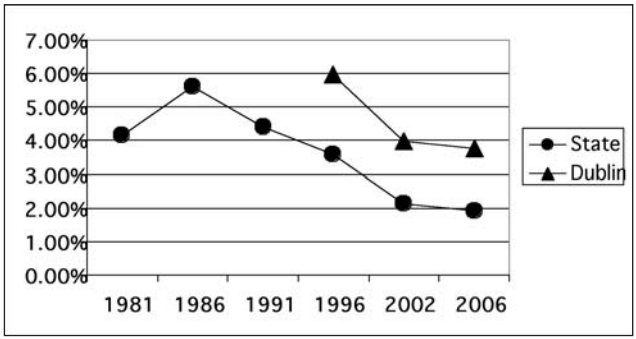
[1] Leon Trotsky (1879-1940) - Russian revolutionary

causes for this seemingly terminal decline will be examined and contrasted with the experience in other countries to see if there is any hope left for the bicycle in Dublin or Ireland.

The Decline

Figure 1 clearly shows the decline of the bicycle in Ireland. Every year since 1986 the proportion of people using a bicycle to commute has fallen and the trend seems destined to continue. Conversely, this decline has been met by a huge increase in people driving to work. In 1981, 44.1% of people used a car to commute; by 2006 the figure had increased to 62.6% (CSO, 2006). In a country nowhere near its target to reduce greenhouse gas emissions under the Kyoto protocol and with rising levels of obesity, the continued switch from bicycle to car has serious implications.

Figure 1. Percentage Travel to Work by Bicycle[2]



The Causes

There are several reasons for the decline of cycling in both Dublin and Ireland. The chief causes of this decline, including poor transport and housing planning and changing consumer preferences will now be examined.

Urban Sprawl

Many contributing factors to the decline of cycling in Dublin are linked to

[2] Source: Census online at www.cso.ie

planning; be it housing, land use or transport planning. Poorly planned housing development has allowed the greater Dublin region to expand into neighbouring counties. It appears the dubious doctrine of 'any development is good development' has been slow to die off. High house prices combined with a lack of suitable high density accommodation and personal housing preferences, have seen Dublin expand outwards rather than upwards (EEA Report, 2006). This has led to a situation in which many commuters are living on the far side of the M50. Without adequate public transport links, suburban dwellers must resort to their cars.

People are most likely to use bicycles for short journeys and only a small proportion are willing to cycle over a distance of 10 kilometres.[3] A study on cycling in Dublin showed that distance was the fifth largest reason for motorists not cycling, with 16% citing this reason (Keegan and Galbraith, 2005). Clearly Dublin's urban sprawl is a major impediment to cycle commuting, and with a higher proportion living further and further away from the city it is natural to expect a decrease in the overall proportion of commuters cycling.

Transport Planning: Moving in the Wrong Direction

A government minister, Mary Harney, once noted that Ireland is 'closer to Boston than Berlin' and this is especially true for transport planning. Ireland has unfortunately followed the United States' lead with its penchant for auto-centric planning. The motorisation of cities has made them less attractive for walkers and cyclists, with planners viewing the city from 'behind a windscreen'. Planning policy has gravitated towards catering for cars with the result that most city centre streets are merely motoring throughways, with pedestrians or cyclists perceived as an 'obstacle' to traffic flow. Solving traffic congestion by building more roads is the equivalent of digging oneself out of a hole. Road pricing and discouraging car usage in city centres through restrictive measures such as traffic calming are both sensible options. However, such policies remain political hot potatoes that few are brave enough to handle. Car orientated policies should not be underestimated in the deleterious effect they have on bicycle usage; one econometric study found that such policies accounted for 40% of the differences in bicycle usage levels (European Cyclists Federation, 2006). It is unfortunate that we have taken our lead from the United States' unsustainable model instead of looking to Europe, where cities like Amsterdam and Enschede in the Netherlands have planned around the needs of pedestrians and cyclists. The following sentiment from a Dutch councillor in Enschede highlights a belief

[3] According to the 2006 census figures only 10.3% of cyclists travelled over 10km to work.

not held in Ireland: 'our city's main infrastructure does not only belong to cars, it is everybody's' (European Cyclists Federation, 2006).

Cycle lanes in Dublin are of notoriously bad design without any regard for proper continuity or integration of the system. Cycle lanes in many cases are there just to pay lip service to cyclists, not to actually address their practical needs. Evidence of this can be seen in the fact that cars are legally allowed to park on some cycle lanes in the evenings and at weekends, while cycle lanes often end abruptly onto a busy road. The continued lack of inclusiveness of non-motorised transport modes in Irish planning is one factor contributing to cycling's decline in this country.

Attitudes: Crowding out the Cyclist

Policy can lead the way in changing attitudes. If good policies are established the public will most often follow them. An example of this includes the Irish smoking ban, which despite the initial outcry has been a complete success. Another example of policy changing attitudes was the 'zero tolerance' policy on crime in New York which contributed to a hugely diminished crime rate. The same principles could and should be applied to transport policy and the attitudes it fosters.

The application of 'signalling theory' is as relevant here as it is to principal-agents' reactions to central bank announcements. If the agent (in this case the government) makes an announcement, the principals (the citizens) will react to it in a way that benefits them the most. So if the agent's priority is to focus on road and motorway building, that sends a signal to the principals to use their cars more. If the agent announces urban traffic calming measures and road pricing, the principals will take this information on board and reduce their car usage accordingly. Constantly increasing the traffic volume capability of roads sends the signal that cars should be used as the main form of personal transport. Dublin city centre is of medieval design but has been adapted for motorists and for use as a throughway for heavy trucks making their way from Dublin Port. As a result pedestrians and cyclists have been sidelined and overlooked as valid road users. The opinion amongst many is that the streets are far too dangerous for cycling (Keegan and Galbraith, 2005; Irish Times, 2007). The motorisation of the city and decreasing safety for non-motorised modes of transport is another major reason for the decline of cyclists.

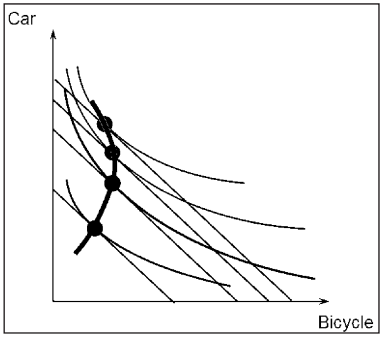
The Bicycle: An Inferior Good

In Ireland the bicycle has been transformed from a normal good to an inferior good. While there are no available statistics for Irish bicycle usage fifty years ago, it is certain that a much larger proportion of people used a bicycle as their

main form of transport. UK figures from the 1950's show the bicycle was used in 15% of all trips and the Irish figure was likely much higher (Pucher and Bueher, 2008). The following quote from a bicycle shop owner gives an idea of past usage and the attitudinal change that has taken place: 'Nearly everyone had a bike then, doctor or tailor…there were lots of women cyclists in the 1950's. All changed now, all the young girls want cars these days' (Irish Times, 2006). Further evidence for the transformation of the bicycle to an inferior good can be found in a survey conducted in Dublin (Keegand and Galbraith, 2005). In this survey only 21% of car commuters would consider cycling to work, with the largest reason cited for not cycling being 'I prefer to drive' (ibid). This does provide strong evidence for the assertion that the bicycle has transformed from a normal good into an inferior good in the mind of consumers. Changing preferences are likely linked to changing economic conditions. Bicycles are seen as a substitute for cars rather than as a possible complement. Therefore they are not seen as desirable, as only people who cannot afford a car use this 'inferior' substitute. This mindset was concisely captured by a past deputy mayor of Shanghai, China: 'the bicycle is just a reminder of past poverty' (Hilary, 1997).

In contrast it is interesting to note that in the Netherlands, Germany and Denmark cycling is distributed evenly across all income groups. This suggests that the bicycle is a normal good in these countries, as rising income does not affect bicycle usage (Pucher and Bueher, 2008). The bicycle's income inelasticity in those countries may be attributed to social norms as cycling is perceived as 'normal' behaviour for everybody to engage in, not just for poor people, children or athletes. The bicycle is seen as a complement to the car, as it is used for short journeys and the car for longer ones. Figure 2 below illustrates Irish consumer preferences in relation to bicycles and cars.

Figure 2. Income Offer Curve for Bicycles and cars in Ireland



The experience in other countries

Countries that experience high rates of economic growth that results in transformation from a poor to a wealthy economy over a short time period are likely to have low levels of bicycle usage. The private car is seen as a symbol of success, and myopic transport policy centres on building more and more roads to accommodate the increasing number of cars. However there is a cohort of mature economies that take a more enlightened and holistic approach to transport planning and policy. Some of the best examples are the Netherlands, Denmark and Sweden. In these counties bicycle usage declined to low levels in the 1970s but a reversal of transport and urban planning policies revived cycling to its present healthy state.

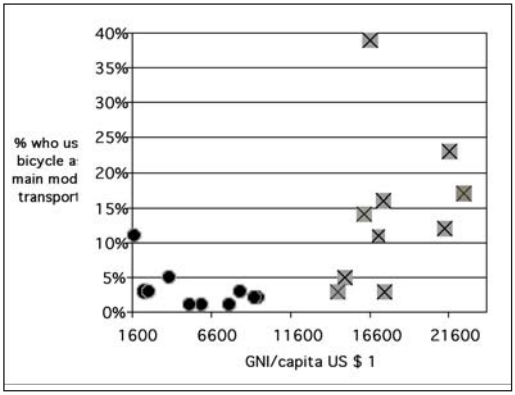
The construction of Figure 3 gives an informal insight into the difference in levels of cycling uptake between countries that have been developed for a long time and those that have only recently achieved similar levels of development.[4] While it is not a large sample, it gives a rough idea of the trends among different European countries. The countries are divided into two groups; wealthy and poor, with wealthy being represented by square markers, poor by circular ones.[5] A country is defined as wealthy if its 1988 GNI surpassed $10,000. The year 1988 is used, as a country's policies would have a lagged effect on the behaviour of its citizens, especially with regard to an issue such as cycling which requires an attitudinal change. As such, 20 years should be enough time to produce an effect. One might expect that a poorer country would have more people using bicycles than a richer one, but it is also likely to have poorer planning. Many countries that were considered poor 20 years ago are now wealthy, with motorisation of the masses often taking priority over other forms of transport. Countries such as Ireland that were coming from a poor background would likely have viewed the bicycle as a reminder of backwardness and the car as a symbol of progress. Overlooking the bicycle as a valid means of transport both in policy and road design is a sure way to displace cyclists from the roads. A good contemporary example of the transition from poor to rich economy and its detrimental effect on cycling can be seen in Beijing today, where cycle lanes are being appropriated for the use of cars (China Daily, 2004).

---

[4] Figure 3 was constructed using data from the UN statistics division and the Flash Eurobarometer (2007) article 'Attitudes on Issues Related to EU Transport Policy' published by the European Commission.

[5] Wealthy countries include: Austria, Belgium, Denmark, Finland, France, Germany, Italy, Netherlands, Sweden and the UK. Poor countries include: Bulgaria, Cyprus, Czech Republic, Greece, Ireland, Malta, Poland and Portugal.

Figure 3. GNI/capita 1988 and % using Bicycle as Main Mode of Transport 2007



A rough trend can be identified in this simple graph showing that the countries that have been wealthy since at least the 1980s have a higher proportion of people who use a bicycle as their main form of transport. The probable reason for the visible disparity in the levels of cycling among the two groups is due to a country's policies and resultant attitudes. Transport planning that takes cyclists into consideration is more likely to engender better attitudes towards cycling than auto-centric planning. In most cases a country that is wealthier will have a more enlightened approach to planning and development than a country that is poorer. Countries with well established economies tend to give greater consideration to ideas of sustainability, the environment and urban gentrification, while poorer countries are more concerned with rapid economic development, leaving such issues take a back seat. As the graph suggests, bicycle usage ultimately comes down to planning and resultant attitudes. It is no surprise that the Netherlands, a country that actively pursues pro cycling policies, has the highest level of bicycle usage at approximately 40%. Positive policies foster positive attitudes towards cycling, meaning that it is seen as a normal and valid mode of transport.

In poorer countries the car is seen as much more of a status symbol, as status symbols must be scarce to confer status upon their owners, whilst in a wealthier country of material abundance cars hold less cachet (Blumberg, 1974). This poor man's mentality may still exist in countries that have grown wealthy since the eighties (for instance Ireland). Roads congested with cars are seen as indicative of economic success.

Solutions

The decline in cycling in Dublin and Ireland may appear to be terminal but it is not. There is much room for improvement. Taking the cities of the Netherlands as a guide will do a lot to help inform future policy addressed at tackling the issue. A two pronged approach focused on policy and changing attitudes through education will achieve the most favourable results. Practical policies to pursue include; the construction of continuous and safe bicycle lanes, traffic calming such as lower speed limits and physical obstacles for cars, de-motorisation of city centres, adequate bike parking facilities, training and education, revised traffic laws and integration with public transport. These policies have proved successful in the cities to which they were applied in the Netherlands (Pucher and Bueher, 2008).

However, Ireland has a long way to go to fulfil such recommendations. For Ireland a change in thinking amongst all levels of policy makers as well as an attitudinal change amongst the public are required if we want to keep the bicycle on our roads.

Bibliography

Blumberg, P. 1974. 'The Decline and Fall of the Status Symbol: Some Thoughts on Status in a Post-Industrial Society'. *Social Problems* 21:4:480-498.

China Daily. 2004. *China ends 'bicycle kingdom' as embracing cars*. 11/11/2004.

European Environmental Agency Report. 2006. 'Urban sprawl in Europe: The ignored challenge'. No.10/2006. Viewed at: http://reports.eea.europa.eu/eea_report_2006_10/en/eea_report_10_2006.pdf

European Cyclists Federation. 2006. 'Continuous and Integral: The Cycling Policies of Groningen and Other European Cycling Cities'. *Fietsberaad* Publication 7. Viewed at: http://www.ecf.com/files/2/12/22/060608_Fietsberaad.pdf

Flash Eurobarometer 206b - The Gallup Organisation. 2007. 'Attitudes on issues related to EU transport policy'. Viewed at: http://www.ecf.com/files/2/12/22/070814_EU_mobility_attitudes_survey.pdf

GNI data from UN statistics division:
http://unstats.un.org/unsd/snaama/dnllist.asp

Gordon, D. 1991. *Steering a New Course; Transportation, Energy and the Environment*. Washington DC: Island Press.

Hilary, J. 1997. 'Paradise lost: The decline of bicycle transportation in Asia' International Bicycle Fund. Viewed at:
http://www.ibike.org/economics/asia-bicycle-decline.htm

Irish Census Data, Central Statistics Office. Viewed at:
www.cso.ie

The Irish Times. 2006. *Dundrum Bicycle Shop that's Stayed the Course*. 27/9/2006.

The Irish Times. 2007. *Pedestrians account for 50% of fatalities on Dublin's roads*. 09/06/2007.

Keegan O.P. and Galbraith, J. 2005. 'Attitudes of cyclists and car commuters to cycling in Dublin'. Viewed at:
http://home.connect.ie/dcc/docs/DublinCity/Attitudes.pdf

Pucher, J. and Bueher, H. 2008. 'Making cycling irresistible, lessons from the Netherlands, Denmark and Germany'. *Transport Reviews* Vol.28. Viewed at: http://www.policy.rutgers.edu/faculty/pucher/Irresistible.pdf

TAKING FLIGHT: THE ECONOMIC CONSQUENCES OF THE EU-US OPEN SKIES AGREEMENT

EMMA KEARNS

Senior Sophister

We have grown accustomed to low-cost flying within Europe thanks to recent efforts to liberalize what was historically a heavily regulated industry; however, the first agreement pertaining to transatlantic air travel between the EU and the US was signed only as recently as April 2007. Emma Kearns investigates the impact of the 'open skies' agreement on the aviation industry, giving particular mention to its predicted knock-on effects for Ireland. Increased transatlantic competition is expected to benefit consumers through lower prices, as well as improved choice and quality of service. Non-passenger considerations and the implications of issues such as the environment and the looming US recession are also discussed.

Introduction

The 'open skies' agreement is the first aviation treaty signed by the United States and the European Union. It is due to take effect from the 30th March 2008. The primary objective of 'open skies' or an Open Aviation Area (hereafter OAA) is to eliminate the restrictions that are currently implemented through the traditional bilateral air service agreements. It will remove constraints on the destination, frequency of flights and on the number of carriers authorized to fly between the two regions:

> 'The ultimate objective of the European Union is to create an Open Aviation Area: a single air transport market between the EU and the US in which investment can flow freely and in which European and US airlines can provide air services without any restriction, including access to the domestic markets of both parties' (European Union: Delegation of the European Commission to the USA website, 2007).

This preliminary agreement paves the way for the aviation industry to be regarded as any other 'normal' global industry. While this is the first stage agreement between the parties it is expected to act as a blueprint for world aviation.

This paper will briefly discuss the history of the aviation industry and how it has evolved from a highly regulated market to the current industry structure. It will also investigate the implications of this agreement on market operators and consumers, giving a concise analysis of non-passenger concerns such as employment, cargo and airport slots. Special mention is also given to the implications the agreement may have on aviation in Ireland. Furthermore, consideration will be given to environmental issues, the fragile US economy and the potential impact of a recession on the aviation industry.

A Brief History

Traditionally the aviation industry has been characterised by protectionist policies, resulting in high prices, low productivity and inefficiency in the market. Typically the state owned airlines or 'flag carriers' dominated the market and as a result enjoyed monopoly power. A key feature of a monopoly is barriers to entry which prevents other firms entering the market. A monopolist has the ability to set price above marginal cost. This gives rise to the existence of supernormal profits in both the short- and long run. Owing to the fact that a monopoly does not operate at a price equal to marginal cost they are inefficient and generally wasteful of resources. Notably, incumbent airlines strategically used regulatory capture to keep new entrants out of the market.

The US were first to deregulate the aviation industry with the Airline Deregulation Act of 1978. In Ireland the market was deregulated through the Anglo-Irish agreement in 1986 which opened up the skies between Ireland and the UK. On the first day that deregulation was introduced airfares on the Dublin-London route fell by 54% (Barrett, 1997). This remarkable fall in prices illustrates the positive effect deregulation can have on a market. By allowing competition into a market it forces incumbents to lower prices or risk losing their market share to new entrants. Deregulation in the rest of Europe followed in 1993 with the creation of the Single European Market, with the policy of protectionism ending on April 1st 1997. Although the aviation sector has clearly come a long way, restrictions still remain on airlines operating between the US and the EU. The objective of the new 'open skies' agreement is to remove the barriers that currently restrict competition.

Market Structure

At present the transatlantic aviation market is not perfectly competitive. A perfectly competitive market is characterised by a large number of firms, no barriers to entry or exit and, firms operating at a price equal to marginal cost. This implies that no supernormal profits are earned. In practice a perfectly competitive market is difficult to achieve. Baumol's theory of contestable markets serves as a more accurate approximation of reality. It does not require a large number of firms in an industry as is the case in perfect competition. However, it yields the same competitive results: 'A contestable market is one into which entry is absolutely free, and exit is absolutely costless' (Baumol, 1892: 3). It is the threat of new entrants that confers contestability to a market. By eliminating some of the barriers that hinder competition in the aviation industry, the OAA is a step towards full liberalisation.

The OAA will increase competition for transatlantic carriers. This will compel the existing incumbents to operate more efficiently or face significant losses. It will also reduce their relative market power as more firms are authorized to enter the market. The liberalisation of the industry will enable firms to expand their capacity utilisation through new and increased services. This will facilitate the move towards an optimal level of production i.e. at the lowest point on the average cost curve. As all carriers in the industry converge towards optimality, it will augment the efficiency of the entire market, not just on the transatlantic routes. Airlines will be able to pass on cost savings to consumers within the EU and US regional markets.

Deregulation will generate economies of scope through increased network size, possible mergers and alliances. It will allow carriers to operate on a larger scale without hindering the competitiveness of the market. Firms will be able to coordinate their activities in order to reduce costs and subsequently pass on the saving to customers in the form of lower fares. Consolidation appears to be the way forward. Prior to the implementation of the 'open skies' agreement airline mergers were already in progress. Recent mergers include Air France and KLM, Easyjet and GB Airways and Lufthansa and Swiss Air. These mergers have the potential to exploit scale economies in allowing:

> 'Two carriers to spread certain fixed costs over more passengers (scale economy). The carriers might achieve added savings by reconfiguring their combined network to connect more flights to certain hub airports (scope economy). They might also achieve higher utilisation e.g., by combining traffic to raise load factors (density economy)' (The Brattle Group, 2002: 5).

These findings demonstrate the probable cost savings that can be accomplished. Through an OAA, overall efficiency of the market can be enhanced.

Consumers

Increased competition will lead to lower fares for passengers. This in turn will stimulate demand in the transatlantic market. It is estimated that the agreement will generate twenty-six million additional passengers over the first five years (Booz Allen Hamilton Ltd., 2007). The rise in passenger numbers will provide stable growth for the transatlantic market. Lower fares will increase consumer surplus. The magnitude of the growth will be determined by the price elasticity of demand.

Figure 1. Consumer Surplus Diagram[1]

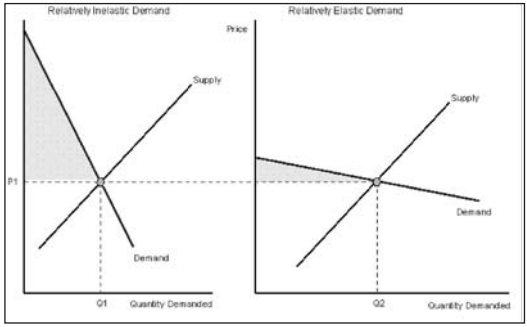


Figure 1 illustrates that the more inelastic the demand for a commodity the higher the potential consumer surplus as some people will be willing to pay a high price in order to continue consuming the product.[2]

The estimated consumer surplus over the first five years of the agreement is shown in Table 1. It demonstrates that consumer surplus is expected to lie between €6.4 billion and €12 billion depending on how responsive consumers are to changes in price. The greater the sensitivity to price, the more impact a

[1] Source: http://tutor2u.net/economics/revision-notes/as-markets-consumer-surplus.html

[2] A consumer with relatively elastic demand is more responsive to changes in price, thus a price increase will cause a more than proportionate fall in quantity demanded, as consumers will switch to close substitutes. In the case of relatively inelastic demand consumers are less responsive to changes in price. A rise in price in this case will cause a less than proportionate fall in quantity demanded.

price reduction will have.

Table 1. Estimated Consumer Surplus[3]

Year	Price Elasticity = 1	Price Elasticity = 2
1	1,000	410
2	2,000	850
3	2,800	1,300
4	3,200	1,700
5	3,400	2,100
Total	12,000	6,400

In addition to this gain, consumers will also benefit from the 'open skies' agreement through improved choice and quality of service. As the number of carriers grows, firms will endeavour to improve their services in an attempt to gain market share.

Non-passenger Considerations

Employment

The benefits of the 'open skies' agreement extends beyond that of the consumer. One of the consequences of an enlarged transatlantic market will be the creation of additional employment:

> 'The elimination of output-restricting bilateral agreements is estimated to range from 2,800 to 9,000 employees in the EU aviation industry. The increased employment in the United States from both effects is estimated to range from 2,000 to 7,300 additional employees' (The Brattle Group, 2002: 75).

Indirect employment will also be augmented in supporting industries.

Cargo

The air cargo industry is a sizeable market. Through the liberalisation of the aviation industry it will provide this sector with an impetus for expansion. At present air trade accounts for 50% of total merchandise traded between the EU and US. Growth of between 1-2% of the existing market is expected (Booz Allen

[3] Source: Booz Allen Hamilton (2007: 7)

Hamilton, 2007). The enlargement of the cargo industry will in turn generate auxiliary employment in this sector.

Airport Slots

> 'Airport slots or units of capacity at capacity-constrained airports are allocated to airlines in order of seniority: that is according to an airline's grandfather rights at the airport' (Barrett, 2007: 64).

This method of slot allocation hinders competitiveness at hub airports. The demand for these slots will be significantly enhanced through an OAA as airlines launch new routes in response to the new emerging market. At present the price of slots in hubs are expensive. Where capacity is constrained and demand continues to escalate, the cost is expected to increase further. In 2006 slots at Heathrow cost €15million per pair (Barret, 2007). Carriers face an opportunity cost regarding the retention or sale of these slots. There has been a shift towards the use of outlying airports where there is spare capacity. Ryanair have taken advantage of such airports and consequently reduced their cost base. The airport charge for Aer Lingus was shown to be 85% larger than that of Ryanair in 2007 (Barret, 2007). Owing to the fact that Ryanair fly into smaller airports it allows them to achieve complete turnaround in twenty minutes. This efficiency could not be accomplished at a hub airport. This illustrates the potential savings that could be made if carriers substitute away from hubs. This may become a viable strategy for many carriers as the traffic at hubs is expected to be augmented through an OAA.

What Does 'open skies' Mean for Ireland?

As discussed above, the 'open skies' agreement between the EU and US will liberalise the transatlantic aviation market. Consequently the market will become more competitive as the barriers that previously restricted entry will be eradicated. With the intention of gaining an understanding of how this agreement will affect the Irish economy, a reflection on the influence deregulation initially had is required. Deregulation had a profound influence on the Irish aviation industry. It resulted in a remarkable reduction in airfares with fares on the Dublin-London route more than halving on the first day of deregulation. Lower airfares were not the only consequence however; the tourism sector also boomed:

> 'In 1994 the Irish Government Green Paper on Aviation Policy estimated that airline deregulation had generated a 60% increase in visitor numbers, additional tourist earnings of £560m and an additional 25,000 jobs in tourism over the years 1987-1993' (Barrett, 2007: 71).

Deregulation enhanced access and enabled tourists to travel to Ireland more affordably.

With this in mind, the 'open skies' agreement ought to have a positive effect on the Irish economy in terms of employment opportunities, tourism and lower fares for consumers both in the transatlantic and regional markets. Ireland is a FDI intensive country with most of the investment coming from the US. By 2006 Ireland had attracted 470 firms from the US alone, creating 95,515 jobs (IDA Ireland website). This is a considerable market and through an OAA greater access to Ireland via the US is possible. Perhaps this will further stimulate FDI investment into Ireland in the coming years.

The regional airports in the West of Ireland will be net beneficiaries of an OAA. Table 2 illustrates the number of routes served by the nine airports of the western region. This is comparable to the 150 routes served by Dublin airport.

Table 2. Number of Routes Served from Irish Atlantic Region Airports, Summer 2007[4]

Shannon	47
Cork	40
Knock	18
Galway	16
Derry	12
Kerry	6
Waterford	5
Donegal	2
Sligo	2
Total	**148**

As previously noted, when airlines opt for outlying airports there is substantial time and cost savings to be made. It is evident from the figures above that these airports could serve as an alternative to Dublin. Airlines that intend to take advantage of the liberalised transatlantic market have the option of using

[4] Source: Barrett (2007: 69)

these regional airports to facilitate the addition of new routes. They also have the capacity for use as a base by US carriers. These regional airports are likely to benefit in terms of employment due to increased passenger traffic.

Further Considerations

Environmental concerns

As the 'open skies' agreement is expected to boost the volume of air traffic it will have auxiliary implications for the levels of pollution. The aviation industry currently constitutes 3% of greenhouse gas emissions (European Commission, 2008). This figure is set to rise due to the expected escalation in worldwide travel. Air travel also contributes to the levels noise pollution. As a result, the EU has launched a Clean Sky initiative:

> 'By 2020, Clean Sky hopes to cut emissions of carbon dioxide by 50%, nitrogen oxide by 80% and noise pollution by 50%, as well as setting up an eco-friendly life cycle for products – design, manufacture, maintenance and scrapping / recycling' (European Commission website, 2008).

This initiative aims to diminish the pollution caused by the aviation industry.

World Economy

The state of the world economy is liable to have an influence on the aviation industry. The US economy is on the brink of a recession. The subprime crisis, combined with a growing budget deficit and weak dollar, has left the US economy in a vulnerable position. The weak dollar in relation to the euro renders European goods more expensive relative to US domestic goods. Furthermore, it makes it more expensive for Americans to holiday in Europe. If these exchange rate conditions persist it could reduce the demand for transatlantic flights in the US. Moreover, trade between the US and Europe could diminish, thus impacting on air cargo services.

Due to the network effects of globalisation the subprime crisis has not been confined to the US. Investors and institutions internationally have been affected. If there is a slowdown in the world economy this could have adverse consequences on the entire aviation industry, not just on the transatlantic market. Consumers may opt for fewer journeys abroad if they have less income, thus reducing the demand for both short haul and long haul flights. Moreover, if the price of oil continues to rise, it will inflate operating costs for airlines and

diminish their ability to lower costs. It can be deduced from this that airfares may increase, further depressing demand.

Conclusion

It is probable that the 'open skies' agreement will have positive consequences on the aviation industry as a whole. While this is a preliminary agreement and further issues need to be addressed, it offers the potential to liberalise the transatlantic market. As discussed, deregulation has favourable effects on a market. It compels incumbents to operate more efficiently or be forced out of the market by new entrants. Consumers benefit via lower prices, superior choice and an overall improved service. However with the instability of the world economy at present, it is difficult to fully postulate the true impact this agreement will have. Despite these concerns it is the potential for competition that is imperative for the efficient operation of any market. The 'open skies' agreement will lay the foundations for the future of the entire aviation industry.

Bibliography

Barrett, S. 1997. 'The implications of the Ireland-UK Airline Deregulation for an EU Internal Market'. *Journal of Air Transport Management* 3:2:67-73.

Barrett, S. 2007. 'Hub Airport Slots, Market Exit and Irish Regional Economic Development'. *Quarterly Economic Commentary Winter 2007:* 63-79. Viewed at: http://www.esri.ie/UserFiles/publications/20071220111940/QEC2007Win_SA_Sean%20Barrett.pdf accessed on 04/02/2008

Baumol, W. 1982. 'Contestable Markets: An Uprising in the Theory of Industry Structure'. *The American Economic Review* 72:1:1-15.

Booz Allen Hamilton Ltd. 2007. 'The Economic Impacts of an Open Aviation Area the EU and the US'. Viewed at: http://ec.europa.eu/transport/air_portal/international/pillars/global_partners/doc/us/eu_us_study_executive_summary.pdf accessed on 04/02/2008.

European Commission. 2008. 'Lighter, quieter and smarter – designing tomorrow's planes'.Viewed at: http://ec.europa.eu/news/transport/080205_1_en.htm accessed on 06/02/2008.

European Union: Delegation of the European Commision to the USA. 2007. 'Open Skies: Jacques Barrot in Washington to sign historic aviation deal at the EU-US transatlantic summit'. Viewed at: http://www.eurunion.org/News/press/2007/2007044.htm accessed on 02/02/2008

IDA Ireland Website. 'Foreign Direct Investment in Ireland'. Viewed at: http://www.idaireland.com/home/index.aspx?id=2359 accessed on 05/02/2008.

The Brattle Group. 2002. 'The Economic Impact of an EU-US Open Aviation Area'.Viewed at: http://www.brattle.com/_documents/UploadLibrary/ArticleReport2198.pdf accessed on 06/01/2008

Tutor2u Website. 'Consumer Surplus'.Viewed at: http://tutor2u.net/economics/revision-notes/as-markets-consumer-surplus.html accessed on 05/02/2008.

THE ULSTER CANAL: A COST-BENEFIT ANALYSIS

MARION SHIELS

Senior Sophister

Cost-benefit analysis is a widely used economic policy tool for identifying the efficient supply of public goods. Marion Shiels objectively outlines the estimated costs and benefits of reopening the Ulster Canal, a 'forgotten cultural and heritage jewel'.[1] She identifies the urgent need for investment in the long-overlooked surrounding regions, and points to increased tourism as one of the many benefits associated with this project. With regards to political objectives, a further incentive is the potentially positive role of the rejuvenated canal in promoting greater North-South co-operation.

Introduction

> *'The practice of conservation must spring from a conviction of what is ethically and aesthetically right, as well as what is economically expedient. A thing is right only when it tends to preserve the integrity, stability, and beauty of the community, and the community includes the soil, waters, fauna, and flora, as well as people'.*[2]

The integration of the Northern and Southern waterway systems has recently been of great interest, largely because the potential economic and social benefits are numerous. The reopening of the Ulster Canal would essentially involve the establishment of a wholly integrated inland waterway system, connecting both North and South. Its reopening would be symbolic in establishing North-South links between both communities. The first stage of this project involves the restoration of the Clones-Erne route. The projections are that it will be completed by 2013.

The Irish canal system, which was once a leader in the transportation of

[1] Source: http://www.ulstercanal.org
[2] Aldo Leopold (1887-1948)

heavy-duty goods, now offers a diverse range of economic and social opportunities to the surrounding areas. The supporters of the reopening have highlighted the benefits yielded by similar projects, namely the Shannon-Erne and Belturbet rejuvenation schemes. The main benefits that are expected from the rebuilding of the canal range from economic prosperity due to increased employment and tourism right through to significant social benefits. The potential social benefits include improved scenery, additional services and a potential increase in the currently dwindling population. However, there have also been groups that contest the economic viability and proposed routes of the project.

This case study will analyse the feasibility of reopening the canal by undertaking a Cost-benefit analysis. Firstly, a brief history of the evolution and demise of the Ulster Canal is presented. The key lobby groups behind the reopening of the canal are examined, with a specific focus on the Inland Waterways Association of Ireland (IWAI). This is a voluntary body that focuses on the restoration, rejuvenation and preservation of the waterway systems throughout Ireland. Following this, the potential economic and social benefits of the Ulster Canal are discussed. The town of Clones is focused on, which is planned to be the hub location within the first stage of development. In conjunction with the interviews undertaken with local partnership groups as part of this case study, the potential benefits to this small border town are examined. A review of similar projects which have recently been undertaken questions whether the benefits attained in these projects can be applied to the Ulster Canal case. Finally, a concise and holistic Cost-benefit analysis of the project is given.

The Evolution of the Ulster Canal

The Ulster Canal opened in 1841. It was 92 km in length and stretched from Lough Neagh through an array of towns and villages including: Moy, Milltown, Wattlebridge, Monaghan, Smithbrough and Clones, eventually joining the Shannon-Erne navigation system. The primary purpose of the canal was commercial exploitation and the transportation of heavy-duty goods. However, with the rise of rail and road as transport alternatives, coupled with the inadequate water supply and small locks, the commercial viability of the canal declined in the 1900s. Since its closure in 1931 the canal has fallen into disrepair.

However, it has been widely recognised that the canal could be of significant economic, cultural and recreational benefit to society. Since the 1998 Good Friday Agreement, a number of feasibility studies have been undertaken regarding the viability of reopening the canal (IWAI, 2006). These studies have

found that the proposed costs would reach £90 million. Initially in the first stage, the Irish Government would fund the full capital cost. Operating costs thereafter would be split between North and South:

> 'Reopening of the canal would constitute a large engineering project, which would provide significant inherent benefits in the areas of tourism and economic development. Its strategic value would lie in its contribution to rural development in a disadvantaged area of the border between Northern Ireland and the Republic of Ireland and in its key linkages within Irelands existing 1000 km inland waterway network' (Brady and Mc Cartan, 2005: 4).

Historically the Ulster Canal played a significant role in the economic stability of the towns and villages that surrounded it. However, modes of transport have evolved considerably since the nineteenth century; consequently, we would expect the reopening of the canal to encompass more diverse economic functions. These potential economic and social benefits will be highlighted throughout the course of this essay.

The IWAI has been the key lobby group involved in the rebuilding and rejuvenation of waterways projects throughout Ireland (e.g. the Shannon-Erne project), and has been a leader in bringing forward the proposals for reopening the canal.

Cost-benefit Analysis

The Cost-benefit analysis (CBA) focuses on the region of Clones town and the six miles either side of the Clones-Erne Canal. The economy in these surrounding areas predominately relies on agricultural and regional activities. The tourism industry is extremely primitive and largely underdeveloped. Therefore there are significant grounds for additional development within these regions. The large economic gap that exists between these regions and the rest of the Republic and Northern Ireland could amplify if innovative projects such as the reopening of the canal are not undertaken.

Explaining Cost-benefit Analysis: A Macro-Economic Policy Tool

> 'Cost-benefit analysis is the most important technique for project appraisal in the public sector. In classical microeconomic theory

> the market system leads to maximum efficiency: producers minimise costs, society maximises output, consumers maximise satisfaction and any redistribution of output could not increase one person's satisfaction without reducing that of another' (Mulreany, 2002: 1).

CBA is extremely useful in the analysis of any capital intensive project as a method of measuring the legitimacy of proposed expenditures to expected outcomes. Welfare economics examines the effects of externalities and spillover effects on society. Any CBA must endeavour to include costs and benefits that are not priced by the market system. In order for a project to be implemented, the benefits must be greater than the costs involved. This method of evaluation is particularly important when one attempts to examine the effects of the Ulster Canal on society as a whole.

With projects like the reopening of the canal, monetary gain can be extracted from a wide variety of opportunities such as: boat hire, increased tourism, the renting of docking areas and other recreational activities. From a social welfare perspective it may be extremely difficult to fully measure the effects of the reopening of the canal on the surrounding regions. On closer examination, it is apparent that the majority of the border towns, such as Clones, have suffered economic and social deterioration over the past forty years. Consequently, investment decisions are highly significant for these areas.

However, it must be recognised that both sides of the border have implemented differing methods in the appraisal of capital projects. Capital projects seeking UK Government funding must comply with 'Green Book' guidelines, while those seeking funding from the Republic must follow the Department of Finance Guidelines.[3]

The Costs of Reopening the Canal

Georgi (1973: 18-19) has defined project costs as: 'the value of goods and services that are required to establish, maintain and operate a project'. Capital costs are comprised of land acquisition, construction, and incidental/replacement costs. In relation to construction costs, it was decided that the most efficient route

[3] The differences associated with the two measures relate to the discount rates. The 'Green Book' indicates the use of a 6% discount rate, whereas the Dept. of Finance indicates a 5% discount rate. In addition the inclusions of wider economic and social benefits are not permitted under the 'Green Book' guidelines.

would be to incorporate more river usage, thus bypassing a number of locks that could be transformed into heritage attractions. A study conducted by ESB International and Ferguson McIlveen LLP (2005) estimated the cost of restoration in 2006 prices at £110m capital costs with additional £15m non-capital costs. Another study carried out by Brady and McCartan (2005) also accounts for operational and maintenance costs that would amount to £800,000 per annum.

These figures are broadly in line with other canal rejuvenation projects that have been undertaken in recent years; for example, the Shannon-Erne project had capital expenditure of over £100m and the restoration of the Scottish Lowland Canals in the UK had capital expenditure of £84.5m.

There are also significant costs associated with the employment of canal rangers and personnel to maintain and manage the canal. These costs would be divided between the North and South depending on the proportion of the canal in both areas.

The Benefits of Reopening the Canal

The benefits of a proposed project, as defined by Georgi (1973: 19): 'comprise of all the positive effects, less the negative effects, resulting from the realisation of the project regardless of whom they fall to'. Studies in relation to canal investments have highlighted the central benefits that lie within the economic and social boundaries. The economic benefits include increased employment, new market opportunities, increased tourism, and improved cross-border/cross-community business interactions. The social benefits relate to recreational opportunities, scenic beauty and improved environmental impacts on the district.

> 'The correction of market prices to reflect social rather than market costs and benefits involves the use of shadow prices. McKean lists four ways in which these corrections can be made and the shadow prices derived' (Barrett, 1982: 31).

The four ways in which correction can be made are through: linear pricing, market prices for similar items, governmental choices and the adjustment of market prices to reflect considerations not reflected in these market prices. Consequently, these measures shall be incorporated into the CBA.

Economic Benefits

The economic potential of this region is enormous. The project could open the floodgates for rural and entrepreneurial development in an area that suffers from economic 'fatigue' and a small tourism industry.

Employment:

The areas surrounding the canal are heavily reliant on agricultural and rural industries. With the considerable decline of these industries in other parts of the country, communities in the border regions are experiencing severe economic pressure. The development of the canal offers extensive employment opportunities through a wide variety of industries, including the construction of the canal itself. In a survey undertaken by the IAWI, which questioned 101 businesses along the canal, 40% of employers felt that the reopening of the canal would lead them to hiring new staff. The general consensus that emerged was that considerable support for the reopening exists, with many feeling that the benefits attributed to it outweigh any economic costs.

In terms of the Clones harbour, there are ample possibilities to fully exploit the opportunities available. It has been considered that an investment in a marine area would be of great economic significance as there are no such amenities in close proximity along the Erne and Shannon-Erne Canals. Thus there is a potential market opportunity for the provision of amenities and services for the people utilising the marina.[4]

Other prospects for economic development includes development of accommodation and hospitality facilities, increased provision for leisure activities, housing development, and business opportunities in the form of land and water based activities.

Increased Tourism:

There has been substantial growth in the tourism industry throughout Ireland since the beginning of the Celtic Tiger and this trend is expected to continue. There is huge potential to improve the ailing tourism industry that exists along the canal's banks. Inland waterway systems offer the opportunity to promote walking, cycling, canoeing and canal trips along the newly interconnected waterways systems. Inland waterways and the associated recreational activities are a potential tourist attraction. With the success of the Shannon-Erne Waterway in providing and promoting a successful tourism promotion plan, the

[4] The average cost per annum for boats in marinas range from £2000-£3000. Thus if there are 30 boats per marina, the total return in investment is £60,000 per annum. Thus it seems that this is an extremely attractive investment.

future for the Ulster Canal looks bright.

Indeed, in an audit of tourism undertaken for the Monaghan region, it was found that there are considerable economic benefits to be availed of in relation to tourism. It was proposed that the development of wetland in the surround areas, water tourism and the establishment of walking and cycling routes would all boost tourism. The Monaghan region has long been considered a leader in fishing activities, and the reopening of the canal provides the opportunity to advance this industry further. In a study undertaken by the IAWI, it was found that the potential direct expenditure by tourists is estimated to lie between £2.6m-£3.2m per annum. If one allows for the multiplier affect, this rise to £3.1m-£4m per annum.

Revenue From Boating:
The fleets along the Shannon and Erne have been expanding steadily in recent years. Both the hire and private boating sectors have become increasingly popular. Hence, the revenue that could be expected from these fleets is enormous due to the expenditure in local areas, and the potential revenue that could be extracted from locks and harbor usage.[5]

Social Benefits
The social benefits which could be extracted from such an investment project expand well beyond the regions in the immediate vicinity of the canal:

> 'A social valuation is clearly important because of the principal role which recreation plays in canal usage vis-à-vis commercial freight operations. Such a valuation is important in deciding the likely future role of canals for recreation purposes, and whether such national environmental and historical assets ought to be maintained' (Willis and Garrod, 1991: 512).

Regeneration Benefits:
Throughout Ireland and the UK commercial recreation is substantial and includes a diverse range of activities such as: boat hire, fishing, walking and cycling. While there are no direct costs associated with these activities, there are unquestionably numerous benefits. Several methods have been developed to measure the value of these benefits:

[5] The average cost to cross a lock is £1.50 per lock. The Clones-Erne Canal would have 2-3 locks, thus it would be expected that the majority of revenue would be extracted from other canal supporting activities.

> 'The travel-cost-method (TCM), which measures the willingness to pay, consumer surplus and utility, by the link between environmental assets and markets for related private goods using recreational trip expenditures as a proxy for willingness to pay in demand estimation' (Willis and Garrod, 1991: 513).

The potential social benefits that could be extracted from the reopening of the canal also extend to variety of environmental and cultural aspects.

Community Benefits:
The areas surrounding the canal have predominantly declining and ageing populations. With significantly higher levels of unemployment in comparison to other regions in Ireland, they have been left behind regarding investment. The reopening of the canal provides a window of opportunity in which cross-border investments may be exploited and a sense of entrepreneurial hunger can become instilled in the region. Many of the towns located along the canal severely lack the provision of services yet have extensive development potential. The reopening of the canal could provide a cohesive force to bring the community together and promote cross-border trade.

Environmental and Heritage Benefits:
There are also many other social benefits which could be extracted, namely the increased aesthetic value of the region. It may also enhance the local environment by improving irrigation. The project is particularly important from a heritage point of view, as the canal re-establishes a link with the past and reflects a rural Ireland from the nineteenth century.

The canal systems throughout Ireland have proven to be extremely environmentally friendly. Canals feature very low on the list of pollutants in Ireland. Over the last decade, boats that travel along the canal have been forced to install holding tanks and pumping stations have been installed along canal routes. This reduces the amount of effluent being dumped into the canals themselves.

The restoration of the canal is essentially a cross-border investment. There are hopes that it will act as a catalyst in promoting investment both sides of the border which would be extremely beneficial for the community as a whole.

Past Experiences

There have been numerous restorations of inland waterway systems in recent years. In the UK it is estimated that over 2 billion has been invested in such renewal projects. The economic and social benefits attributed to these are considerable.

In Ireland, the most successful and influential of these recent investments lies within the Shannon-Erne system. This was completed in 1994 and has greatly influenced the debate surrounding the reopening of the Ulster Canal. There was no CBA undertaken prior to the construction of the Shannon-Erne route. However, there has been an ex post analysis undertaken. The general consensus is that the project has exceeded all expectations.

Ballinamore is a small town border town located in Leitrim and shares similar characteristics with the town of Clones. It has benefited significantly from the economic spin-off provided by the Shannon-Erne waterway. It has developed a high quality marina, which provides an ideal stopping area along the canal. Its success lies in its excellent facilities that encourages tourist to stay. It has also developed a wide variety of cycle, walking and boating amenities, which have helped strengthen economic activity. There is no reason why the success story of Ballinamore cannot be replicated in towns such as Clones. In many ways the redevelopment of the Ulster Canal could use the experience of the Shannon-Erne project as a benchmark. Thus, if one is to take Ballinamore as a prime example of the benefits that could might accrue from the redevelopment of the Ulster Canal, the future looks bright for areas along the canal.

Summary of CBA

The principle aim of a Cost-benefit analysis is to gain a greater insight into the potential economic and social benefits that could be achieved through capital investment projects. Under the proposed rebuilding of the Ulster Canal, the general consensus emerging from the feasibility studies and analysis of past canal investments, namely the Shannon-Erne route, is that a rejuvenated Ulster Canal would be of significant economic and social benefit. It is anticipated that the revenues generated through locks and additional canal services will be substantial, and that the expected return on investment to be above the industry average. The total cost of ownership is expected to be met on a yearly basis.

These expectations are based on a number of contributing factors. Firstly, the success of the Shannon-Erne route proved to be influential in the decision to rebuild the Ulster Canal. The situation of the Clones harbour is ideal as there are

a limited number of harbours along the northern end of the Shannon-Erne route. Hence, there is significant potential for economic gains. Secondly, the traffic along the Shannon-Erne and surrounding routes has increased in recent years and is expected to do so for the foreseeable future. This is as a direct result of users transferring from the Grand Canal routes to the Western routes. Overall, the Clones hub has a wide range of aspects in its favour.

It is expected that the regulation of the canal with regards to the development and expansion of the surrounding areas will lie with the local governments. Within the Monaghan region, town planner Paul Clifford has confirmed that the areas surround the Clones region will be developed in a way that maximises the utility and efficiency of the canal for its users. All marketing for the canal will be undertaken by Waterways Ireland. Hence it is expected that the Monaghan region will be effectively promoted, ensuring the success of the project. Despite there being no expectation for commercial activities along the canal, there does seem to be sufficient recreational demand to justify the reopening of the canal within the Monaghan region.

Conclusion

The reopening of the Shannon-Erne waterway in the early 1990's, which linked the Northern and Southern waterway systems for the first time, ignited interest in the reopening of the Ulster Canal. Initially, the restoration would create a canal that stretches from Middletown to Clones, a region that is currently economically and socially disadvantaged.

It has been highlighted that there are considerable economic and social benefits attributable to these areas with the reopening of the Ulster Canal. The area in question has been devoid of any significant investment in recent decades and is in need of sizeable expenditure in order to tackle deteriorating population levels and to regenerate faltering communities. As already discussed within the CBA, the potential benefits from such a project significantly outweigh the associated costs. Therefore, this CBA finds itself in strong agreement with the decision to reopen the Ulster Canal, and would hope that it offers a window of opportunity to border towns like Clones.

Bibliography

Brady, C. and McCartan, M. 2005. *'Middletown to Monaghan: A Social Economic Study of the benefits of the proposed reopening of the Ulster Canal'*. Viewed at: www.crossborder.ie/pubs/ulster_Canal.pdf

Barret, S. 1982. *Transport Policy in Ireland*. Dublin: Irish Management Institution

ESB International and Ferguson McIlveen LLP. 2005. The Ulster Canal Feasibility Study. Commissioned for the Department of Community, Rural and Gaeltacht Affairs.
Viewed at: http://www.dcalni.gov.uk/ulster_canal_feasibility_study_.pdf

Georgi, H. 1973. *'Cost-Benefit Analysis: Public Investment in Transport: a survey'*. London: Butterworth.

IWAI and PWC in asscociation with Tourism Development International. 2006. *'Ulster Canal Socio-Economic Study: Final Report'*. Viewed at:
http://www.iwai.ie/ulstercanal/BRP%20Ulster%20Canal%20Report%20April%202006.pdf

Monaghan Enterprise Board. 2007. 'Audit of Tourism Development Opportunities in County Monaghan 2007-2013. Viewed at:
www.monaghantourism.com/Final%20Draft%20for%20Public%20Consultation%20Sep%2007_1_.pdf

Mulreany, Michael. 2002. *Cost-Benefit Analysis: Readings*. Dublin: Institute of Public Administration.

Willis, K. and Garrod, G. 1991. 'Valuing Open Access Recreation in Inland Waterways: On Site Recreation Surveys and Selection Effects'. *Regional Studies* 26.5:511-524.