Pricing Theory, Financing of International Organisations and Monetary History

Lawrence H. Officer

Routledge Explorations in Economic History

Pricing Theory, Financing of International Organizations and Monetary History

This book presents the lifelong and ongoing research of Lawrence H. Officer in a systematic way. The result is an authoritative treatment of such issues as market structure and economic efficiency where more than one characteristic of a commodity is priced, both in general and in application to shipping conferences; financing of the United Nations and International Monetary Fund; monetary history of the UK and US; and central-bank preferences between gold and dollars.

Pricing Theory, Financing of International Organizations and Monetary History is divided into four core parts; **Pricing Theory** examines multidimensional pricing, defined as pricing when a commodity or service has several characteristics that are priced. Application to oceanic shipping of commodities is emphasized. **Financing of International Organizations** is concerned with country-group conflicts in the United Nations and International Monetary Fund. It is shown which countries have enjoyed relative advantage, and which have suffered relative disadvantage, in financing these organizations. **Monetary History** provides a fresh look at historical experiences of monetary-standard upheavals: the Bank [of England] Restriction Period (1797–1821), the New England colony (1703–49), and the United States (1792–1932). **Gold** considers a crucial time (1958–67), during which central-bank gold-dollar decisions were power-politically determined.

This creative and important new volume will prove to be of great interest to students and researchers involved in understanding the applications of economic theory to the modern world.

Lawrence H. Officer is Professor of Economics at the University of Illinois at Chicago, USA.

Routledge Explorations in Economic History

1 **Economic Ideas and Government Policy**
Contributions to contemporary economic history
Sir Alec Cairncross

2 **The Organization of Labour Markets**
Modernity, culture and governance in Germany, Sweden, Britain and Japan
Bo Stråth

3 **Currency Convertibility**
The gold standard and beyond
Edited by Jorge Braga de Macedo, Barry Eichengreen and Jaime Reis

4 **Britain's Place in the World**
A historical enquiry into import controls 1945–1960
Alan S. Milward and George Brennan

5 **France and the International Economy**
From Vichy to the Treaty of Rome
Frances M. B. Lynch

6 **Monetary Standards and Exchange Rates**
M.C. Marcuzzo, L. Officer, A. Rosselli

7 **Production Efficiency in Domesday England, 1086**
John McDonald

8 **Free Trade and its Reception 1815–1960**
Freedom and trade: Volume I
Edited by Andrew Marrison

9 **Conceiving Companies**
Joint-stock politics in Victorian England
Timothy L. Alborn

10 **The British Industrial Decline Reconsidered**
Edited by Jean-Pierre Dormois and Michael Dintenfass

11 **The Conservatives and Industrial Efficiency, 1951–1964**
Thirteen Wasted Years?
Nick Tiratsoo and Jim Tomlinson

12 **Pacific Centuries**
Pacific and Pacific Rim Economic History since the 16th Century
Edited by Dennis O. Flynn, Lionel Frost and A.J.H. Latham

13 **The Premodern Chinese Economy**
Structural Equilibrium and Capitalist Sterility
Gang Deng

14 **The Role of Banks in Monitoring Firms**
The Case of the Crédit Mobilier
Elisabeth Paulet

15 **Management of the National Debt in the United Kingdom, 1900–1932**
Jeremy Wormell

16 **An Economic History of Sweden**
Lars Magnusson

17 **Freedom and Growth**
The Rise of States and Markets in Europe, 1300–1750
S R Epstein

18 **The Mediterranean Response to Globalization Before 1950**
Sevket Pamuk and Jeffrey G Williamson

19 **Production and Consumption in English Households 1600–1750**
Mark Overton, Jane Whittle, Darron Dean and Andrew Hann

20 **Governance, The State, Regulation and Industrial Relations**
Ian Clark

21 **Early Modern Capitalism**
Economic and Social Change in Europe 1400–1800
Edited by Maarten Prak

22 **An Economic History of London, 1800–1914**
Michael Ball & David Sunderland

23 **The Origins of National Financial Systems**
Alexander Gerschenkron Reconsidered
Edited by Douglas J. Forsyth and Daniel Verdier

24 **The Russian Revolutionary Economy, 1890–1940**
Ideas, Debates and Alternatives
Vincent Barnett

25 **Land Rights, Ethno Nationality and Sovereignty in History**
Edited by Stanley L. Engerman and Jacob Metzer

26 **An Economic History of Film**
Edited by John Sedgwick and Mike Pokorny

27 **The Foreign Exchange Market of London**
Development since 1900
John Atkin

28 **Rethinking Economic Change in India**
Labour and Livelihood
Tirthankar Roy

29 **The Mechanics of Modernity in Europe and East Asia**
The Institutional Origins of Social Change and Stagnation
Erik Ringmar

30 **International Economic Integration in Historical Perspective**
Dennis M. P. McCarthy

31 **Theories of International Trade**
Adam Klug
Edited by Warren Young and Michael Bordo

32 **Classical Trade Protectionism 1815–1914**
Edited by Jean Pierre Dormois and Pedro Lains

33 **Economy and Economics of Ancient Greece**
Takeshi Amemiya

34 **Social Capital, Trust and the Industrial Revolution: 1780–1880**
David Sunderland

35 **Pricing Theory, Financing of International Organizations and Monetary History**
Lawrence H. Officer

Pricing Theory, Financing of International Organizations and Monetary History

Lawrence H. Officer

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

In Memory of Fathers

especially

Joe Officer

Harry Kolichman

First published 2007 by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Simultaneously published in the USA and Canada
by Routledge
270 Madison Ave, New York, NY 10016

Routledge is an imprint of the Taylor & Francis Group, an informa business

This edition published in the Taylor & Francis e-Library, 2007.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
A catalog record for this book has been requested

ISBN 0-203-93403-2 Master e-book ISBN

ISBN10: 0–415–77065–3 (hbk)
ISBN13: 978–0–415–77065–1 (hbk)

Contents

Figures and tables

Figures

Tables

Preface

'Here are collected the croakings of twelve years – the croakings of a Cassandra who could never influence the course of events in time.' So wrote John Maynard Keynes in his Preface to *Essays in Persuasion* (London: Macmillan, 1931). The chapters in the present volume that are reproduced from earlier publications span a time much longer than that of Keynes, 40 years: 1966 to 2005. This period represents almost my entire professional career.

Keynes tried to sway events, to have great impact on public policy. Notwithstanding his disclaimer above, he ultimately succeeded – as few economists before or since did so. The present volume does not represent attempts to influence policy. My goal is, and always has been, to understand how economics manifests itself in the real world. Economic theory, institutional knowledge, and quantitative tools are all applied here, in an effort to achieve this understanding.

Above all, I produced this volume in order to come closer to my own intellectual completion, though not closure – no subject of scholarly pursuit should ever be considered closed – on certain topics. I am perhaps best known in the economics profession for my work on 'purchasing power parity' and 'dollar-sterling gold points', and have devoted more research to these areas than to others. Yet there are a variety of other issues in which I have profound interest. I have published articles and book chapters on these issues, but the work is incomplete. For a long time I have desired to complete this work, or at least, put the work in better perspective or later context.

The result is the present volume. Four themes – excluding 'purchasing power parity' and 'dollar-sterling gold points' – are selected: pricing theory, financing of international organizations, monetary history, and gold. For each theme, which takes the form of the 'parts' of the volume, some of my previously published work is reproduced. Then – as the concluding chapter ('afterword') in each part – that work is extended, supplemented, and related to the work of others.

I thank Robert Langham of Routledge for encouraging me to proceed with this project. Taiba Batool and Terry Clague, also of Routledge, provided excellent help and advice in the process of producing the volume.

The opinions, judgments, conclusions, and other statements in this volume are all my own responsibility.

Lawrence H. Officer
Chicago, Illinois, USA
June 2006

Part I

Pricing theory

1 The optimality of pure competition in the capacity problem*

The problem of optimum ('efficient') pricing under conditions of complementary demands (the 'capacity' problem) has been formulated in its most essential form by Steiner, and the solution has been presented.[1] However, neither Steiner nor the discussants[2] of his article have pointed out that complementary demand functions in no way vitiate the general rule that a purely competitive market structure yields optimum prices. Yet this fact is of substantial interest. It demonstrates the pervasiveness of pure competition as an optimum situation. Joint use of capacity in no way gives rise to a market failure! Indeed it is surprising that Steiner himself did not perceive this because he discusses various pricing schemes which might attain the optimum, even without knowledge of demand conditions on the part of the price-maker. He concludes that such measures cannot be guaranteed to achieve the efficient prices.

The purpose of the present note is to demonstrate the success of one measure, on the proviso that Steiner's implicit assumption of monopoly in supply is dropped. The conventional theorem that 'under constant costs pure competition maximizes consumers surplus subject to the constraint that producers suffer no monetary loss' is applicable to the capacity problem, at least in its Steiner version. Indeed this constrained maximization is Steiner's criterion of optimality.

The argument applies equally to the 'firm peak' and 'shifting peak' cases. Consider Figure 1.1 (Figure I in Steiner's article).

D_1 and D_2 are the demand curves for capacity and β is the (joint) unit cost of capacity. D_c is the vertical sum of D_1 and D_2. Prices are defined net of operating costs. Optimum prices are: $P_1 = \beta$, $P_2 = 0$ in the firm peak case, $P_1 = \overline{P}_1$, $P_2 = \overline{P}_2$ in the shifting peak case.[3]

It may be wondered why the demand functions D_1 and D_2 are designated as 'complementary.' Actually, no complementarity as usually defined is involved. Each demand depends only on its own price, and not on the other price. Demands are complementary or joint with respect to supply, not from the point of view of the consumers. In calling such demands complementary I am continuing a solecism apparently started by Steiner.[4]

Now suppose that the market structure is one of pure competition. What

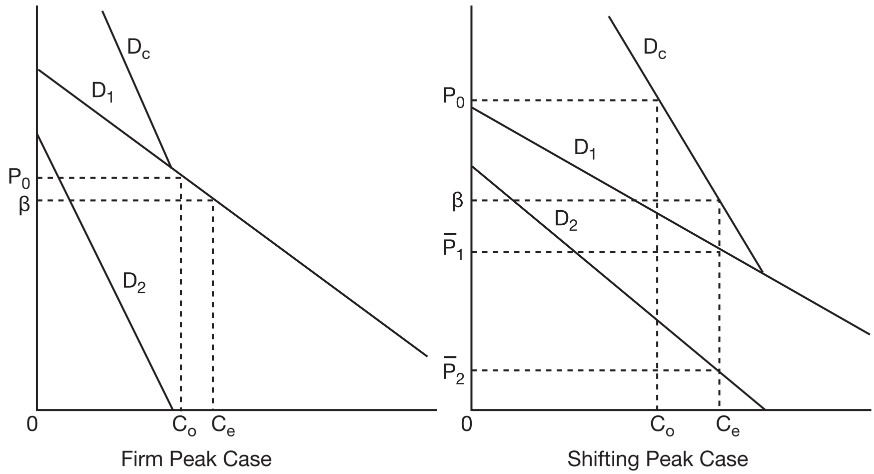

Figure 1.1 Firm-peak case and shifting-peak case: continuous capacity.

does pure competition imply for the behavior of suppliers of capacity? Each firm acts as if it cannot affect prices (P_1 and P_2) and firms act *alone*, collusion and cartellization being incompatible with pure competition. Therefore each firm offers its existing capacity to the limit at the going rates, if it offers capacity at all, i.e., if it is in the market. Firms do this for each of the complementary markets. If total revenue is greater than or equal to total cost, firms offer all their capacity. If total cost is less than total revenue, profits are negative and some firms (plants) will leave the market (capacity will decrease). There is also free entry (expansion of capacity) into the market, which will occur if total revenue exceeds total cost, that is, if observed profits are positive.

Suppose C_o is the amount of capacity in the market. Then a profit of $(P_o - \beta)$ is earned per unit of capacity. The number of firms (plants) increases until capacity equals C_e. Beyond C_e a loss per unit of capacity is incurred and the number of firms (plants) decreases until C_e is reached. Therefore the efficient prices are obtained.

It is obvious that the effective value of β is constant over all firms. The reason for this procedure is that it exactly corresponds to Steiner's assumption of a unitary homogeneous production function (β independent of output). Such a production function under monopoly is precisely analogous to the production conditions of a purely competitive industry of *identical* firms, irrespective of the actual technology of the firms.[5] Were it to be assumed that firms *differ* in efficiency, this would correspond to a monopolist that faces *decreasing* rather than constant returns to scale.

The nature of supply under pure competition requires re-emphasis. The amount of capacity supplied or offered is identical in the two markets and is offered to the limit irrespective of prices. This is the meaning of price-taking:

going rates are accepted. Changes in supply take the form of bodily shifting this inelastic offering. In other words, the supply curve is vertical at the same level of capacity in each market. Shifts in supply merely change the common level of capacity. The outcome of these properties of supply is that the demand functions are summed vertically (as required by the optimality criterion) to create an aggregate demand function. The results of this addition prevail even if the construction is not explicitly done.

There is an aura of unreality to all this because of the acceptance of Steiner's assumption that the cost of capacity is given by a function whose range is *continuous*. This means that to achieve the optimum capacity C_e it must be assumed either that each plant has infinitesimal capacity or that the capacity limit of a plant is continuously variable over a finite interval.

What if the assumption of constant costs were retained but that of continuity dropped? Suppose that a plant provides 0 to C_p units of capacity at a cost of $\beta \cdot C_p$, where C_p, a predetermined number, is identical over all plants. Furthermore, total demand for capacity (in each market) is apportioned equally among all plants – an assumption that is made for convenience in exposition and which does not affect the results. There is no need to assume a one-to-one correspondence between plants and firms, providing parametric (price-taking) behavior is followed by every firm.

If it happens that C_e is divisible by C_p, then the efficient capacity C_e is attained as the equilibrium under pure competition. However, if C_p/C_e is not an integer, then C_e is not attainable under pure competition, nor does it remain the optimum capacity.

The situation is illustrated in Figure 1.2. Capacity is provided by plants each of which is represented by a rectangle with base of length C_p and height of β. The supply of capacity is labeled by the number of plants in the market. First consider efficiency. The capacity C_e is obtainable only by supporting

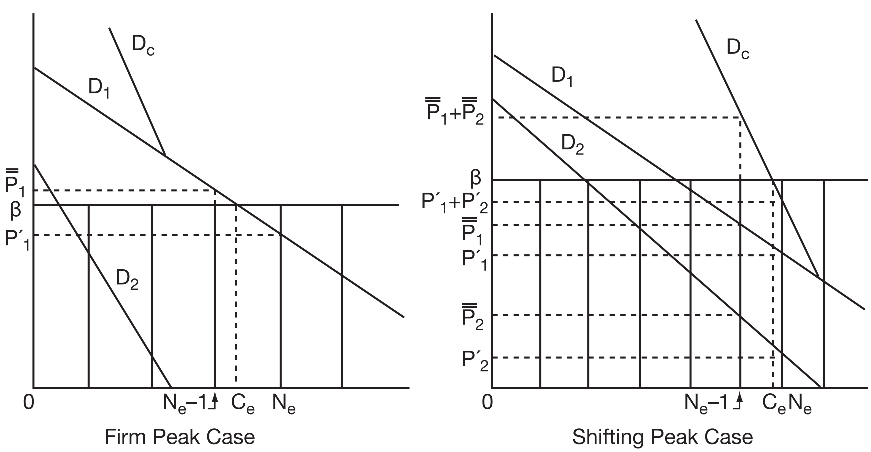

Figure 1.2 Firm-peak case and shifting-peak case; discrete capacity.

(at least) N_e plants, providing $(N_e \cdot C_p - C_e)$ units of excess capacity, and sustaining a monetary loss of $\beta(N_e \cdot C_p - C_e)/N_e$ per plant. (Because of the assumption of equal apportionment of demand, this is not only the average loss but also the actual loss per plant.) The feasible optimum consists of $N_e - 1$ plants, with prices $\overline{\overline{P}}_1$ and $\overline{\overline{P}}_2$ in the shifting peak case and $\overline{\overline{P}}_1$ and 0 in the firm peak case providing this case does not change quality to become the shifting peak case, as is possible. This generates an excess profit per plant of $(\overline{\overline{P}}_1 + \overline{\overline{P}}_2 - \beta)\ C_p$ (in the shifting peak case) or $(\overline{\overline{P}}_1 - \beta)C_p$ (in the unvarying firm peak case). Nevertheless, it is this solution which maximizes consumers surplus subject to no monetary loss on the part of producers.

Does pure competition lead to the efficient solution even in the case of the discrete plant fixed in size? Under purely competitive equilibrium the number of plants cannot be N_e or greater because this entails a monetary loss on the part of producers and hence a decrease in capacity. For example, N_e plants in the market generates a loss per plant of $(\beta - P'_1 - P'_2)\ C_p$ (in the shifting peak case) or $(\beta - P'_1)\ C_p$ (in the firm peak case). If the number of plants is less than $N_e - 1$, positive profits are earned and capacity increases. Once $N_e - 1$ plants are reached, resulting in the feasible optimum, the fact that there remain positive profits will induce additional capacity in the short run. This will result in negative profits and then decreased capacity. Thus oscillatory behavior will occur. However, in the long run it will be understood that the positive profits are due to discontinuity in plant size and become negative upon an increase in the number of plants. Therefore, $N_e - 1$ is the long-run equilibrium number of plants under pure competition, and provides the optimum price structure.

Finally, the case of varying efficiency should be examined. Quite generally, it may be assumed that β differs among firms. Then firms (plants) may be ranked in order of decreasing efficiency, i.e., increasing β. The rectangles in Figure 1.2 will now be of increasing height as they extend from the origin. Clearly, prices are dependent on the capacity cost of the marginal firm (which in general will leave the market). As remarked in the case of continuous capacity, such a situation corresponds to a monopolist with decreasing – not constant – returns to scale, albeit plant indivisibility is now present.

Of course, decreasing returns (monopoly) or, equivalently, varying efficiency (pure competition) implies that optimality must be given the more general definition of the maximization of the sum of consumers and producers surpluses rather than the particular definition stated in the second paragraph of this note. Otherwise there is the problem of expropriating producers surplus (monopoly) or economic rent of intramarginal firms (pure competition) in the form, presumably, of genuine discriminatory pricing.

Notes

* Originally published in *Quarterly Journal of Economics*, vol. 80, no. 4 (November 1966), pp. 647–651.

1 Peter O. Steiner, 'Peak Loads and Efficient Pricing,' *Quarterly Journal of Economics* LXXI (Nov. 1957).

2 Jack Hirshleifer, 'Peak Loads and Efficient Pricing: Comment,' H. S. Houthakker, 'Further Comment,' Peter O. Steiner, 'Reply,' *Quarterly Journal of Economics* LXXII (Aug. 1958).

3 See Steiner, 'Peak Loads and Efficient Pricing,' *op. cit.*, p. 589.

4 'Demands are added vertically because the demands of the separate periods upon capacity are complementary not competitive.' *Ibid.*, p. 588, fn. 8.

5 This fact has been implicitly stressed in quite another context by Jaroslav Vanek. See his *International Trade: Theory and Economic Policy* (Homewood: Irwin, 1962), pp. 187 and 391. Vanek's sole omission is the correspondence between the industry and the single firm.

2 Demand conditions under multidimensional pricing*

1. Introduction

The unidimensional nature of microeconomic analysis has been subjected recently to long-overdue criticism, on the grounds of its distance from reality.[1] It is stressed by critics that a product should be defined not in terms of a one-dimensional unit of measurement but rather in terms of its *attributes, aspects* or *characteristics*, which generally are multidimensional. Nevertheless, this 'abstract-product' approach, to use Baumol's terminology, remains in a primitive state. Advances have been most noteworthy in the theory of consumer demand. Yet even the most elegant studies in this field – those of Baumol [1] and Lancaster [3] – suffer from oversimplified models. Thus Baumol [1, p. 682] must resort to artificial devices to cope with his assumption of completely inelastic demand for any one consumer, and Lancaster makes the strong assumption that the characteristics of a product are identical for all consumers. Quite apart from unrealistic assumptions, however, their models suffer a fundamental deficiency: pricing according to attributes, i.e. *multidimensional pricing*, is not considered.

In the present paper, in contrast, I make quite general assumptions regarding the consumers of a product. Their individual demand functions are downward-sloping. Also, consumers do not behave as one – not only may their demand functions differ but the attributes that each derives from a product may differ. Most important of all, in focusing on *pricing* as a multidimensional phenomenon, I am able to study thoroughly the properties of the demand for a product *expressed in terms of the demand for its attributes*. While for geometrical convenience I consider the two-dimensional case, the results are readily extendible to an arbitrary number of dimensions.

In pricing according to the attributes received by a consumer, it turns out that the price of the product is not uniform over all consumers. Thus it might be objected that I am dealing with a case of discriminatory pricing. However, the variation in price according to the receipt of attributes is *not* genuine discrimination; for the attributes of the product that I consider are measurable and known to the producer as well as the consumer.

One set of situations that fit into my model is covered by the peak-load problem. Provision of public utility services (such as electricity) with peak-time and off-peak-time use as the attributes is the standard example. Another situation is the provision of transportation services, in which the volume and weight specifications of a commodity are the relevant attributes. Still another set of circumstances to which the model applies involves the use of time as an attribute in addition to the conventional dimension of a product. Thus the delivery date or time of performance of service would involve a charge quite in addition to the price of the product itself, with an earlier date presumably having a higher rate. To venture a prediction, I suggest that multidimensional pricing will come to constitute the normal form of pricing in the future, as technological developments at once make products more complicated and facilitate the measurement of their attributes.

2. Assumptions of the model

I consider the market for a certain product, say x, the price of which faced by consumer i is denoted by p_i. Letting x_i represent the number of units of x demanded by consumer, i, his individual demand function is

$$x_i = f_{1i}(p_i) \tag{1}$$

I assume that this demand function is downward-sloping between the axes, that it involves satiety, i.e. $f_{1i}(0)$ is finite, and that a sufficiently high price, say $\tilde{p}_i$, cuts off all demand, i.e. $f_{1i}(p_i) = 0$ for $p_i \geq \tilde{p}_i$.

Product x may be defined in terms of its attributes, say a and b. The number of units of each of these attributes that consumer i obtains per unit of product x is a_i and b_i, respectively. These magnitudes are fixed irrespective of the value of x_i.[2] Define $T_i = b_i/a_i$, the 'trade-off' between the attributes on the part of individual i.

There are a total of n consumers. Thus $i = 1, \ldots, n$. It is assumed that $i \neq j \Rightarrow T_i \neq T_j$. This does not mean that individuals cannot have the same trade-offs. One simply redefines an 'individual' or 'consumer' as the set of all consumers with the same trade-off. The corresponding 'individual' demand function (1) is obtained by adding the basic individual demand functions 'horizontally' (i.e. with price as the parameter) in the usual fashion.

Now, the producers of x follow a multidimensional pricing policy. Let r and q represent the rates per unit of attribute a and b, respectively. Thus the price of product x faced by consumer i is

$$p_i = ra_i + qb_i \tag{2}$$

Though the price of product x varies according to the consumer, the pricing is non-discriminatory *when considered as the pricing of attributes*. Varying the rates r and q according to the consumer would represent genuine discriminatory pricing.

For convenience, I follow the convention that

$$r, q \geq 0 \tag{3}$$

The case of 'dis-attributes' is covered; for if a is a dis-attribute, with, therefore, a negative price r, one redefines a_i as its negative and r switches to a positive sign.

3. The functions to be studied

Combining (1) and (2), we may rewrite the demand function (1) of consumer i as

$$x_i = f_{1i}(ra_i + qb_i) \tag{4}$$

Since a_i and b_i are fixed irrespective of the number of units of x demanded, the function (4) may be re-expressed with the two-dimensional price vector as the independent variable:

$$x_i = f_{2i}(r, q) \tag{5}$$

Proceeding in step, we now alter the dependent variable so that it refers to the *demand for attributes*. With attributes two-dimensional, the dependent variable also is a two-dimensional vector. Letting A_i and B_i denote the demand for attributes a and b on the part of consumer i, his demand function becomes

$$(A_i, B_i) = f_{3i}(r, q) \tag{6}$$

where

$$\left.\begin{aligned} A_i &= a_i f_{2i}(r, q) \\ B_i &= b_i f_{2i}(r, q) \end{aligned}\right\} \tag{7}$$

Letting A and B denote the total demand, i.e. the demand on the part of all n consumers, for attributes a and b, respectively, the market demand function is

$$(A, B) = f_3(r, q) \tag{8}$$

obtained by vectoral addition for each value of (r, q):

$$\begin{aligned} (A, B) &= (\Sigma\, A_i, \Sigma\, B_i) = \Sigma\, (A_i, B_i) \\ &= \Sigma\, f_{3i}(r, q) = f_3(r, q) \end{aligned}$$

The summing of function (6) is analogous to the conventional addition of individual demand curves. The difference is that both demand and price are two-dimensional rather than unidimensional.

The market demand function (8) *is single valued.* To every value of (r, q) corresponds a unique value of (A, B). This is because (8) is obtained by applying addition and scalar multiplication to a set of single-valued functions (5) and the set of all single-valued functions is closed with respect to these operations.

Consider the inverse of the market demand function:

$$(r, q) = f_3^{-1}(A, B) \tag{9}$$

Although (8) is defined for all non-negative values of (r, q), (9) is defined only for those values of (A, B) which satisfy (8). Thus, the domain of (8), hence the range of (9), is the set of all two-dimensional vectors whose elements are non-negative real numbers. However, as will be shown subsequently, the range of (8), hence the domain of (9), is only a subset of this set.

While, as shown above, (8) is a single valued function, (9) is not single valued. In precise terms, (8) is single valued for all points in its domain while (9) is single valued only in part of its domain. However, it will be demonstrated that (9) is 'effectively' single valued in the sense that to each point in its domain corresponds a unique total expenditure, R, on product x. Thus the total-expenditure function

$$R = f_4(A, B) \tag{10}$$

defined as follows:

$$R = rA + qB, \quad \text{subject to (9)}$$

is single valued, as demonstrated finally in Section 6.

4. The case of one consumer

The case of one consumer ($n = 1$) is important not only because it is a first step in the examination of the n-consumer case but also because it represents fully the situation in which the trade-off of the attributes is identical for all consumers, i.e. $T_i = T$ for all i. In the latter situation we can collapse the n consumers to only one, so that the 'individual' demand function *is* the market demand function, as discussed in Section 2 above.

In this section we develop graphical representations of the individual demand function and its inverse. In addition, we show that the expenditure function for the individual consumer, say i, is single valued.

Suppose a given price, say $\bar{p}_i$, is faced by the consumer. According to (1) this produces a unique quantity of product x demanded, say $\bar{x}_i$. To $\bar{x}_i$, in turn, corresponds a unique value of (A_i, B_i), say $(\bar{A}_i, \bar{B}_i)$, given by

$$(\bar{A}_i, \bar{B}_i) = (a_i\bar{x}_i, b_i\bar{x}_i)$$

Now, there exists a set of values of (r, q), say $\{(\bar{r}, \bar{q})\}$, each element of which satisfies both (3) and (from (2))

$$\bar{p}_i = ra_i + qb_i \tag{11}$$

a straight line, which may be rewritten as

$$r = \bar{p}_i/a_i - T_i q \tag{12}$$

The plotted inner straight line in Figure 2.1 is the graph of (12) in the first quadrant, and its set of co-ordinates (including the end-points) is $\{(\bar{r}, \bar{q})\}$. The line is labelled with the unique value of (A_i, B_i) to which it pertains. As the given value of p_i increases, x_i decreases (via (1)) and hence (A_i, B_i) also decreases. Simultaneously, the line (12) moves outward with greater intercepts but the same slope, T_i. Figure 2.1 is everywhere dense with lines of Equation (12) plotted for all non-negative values of p_i.

Two limiting cases may be noted. $p_i = 0$ implies that $r = q = 0$, i.e. $\{(r, q)\} = (0, 0)$, and (12) shrinks to the origin. At that point the maximum x_i and (A_i, B_i) are reached. There is also a limit point in the other direction. For p_i at and above $\tilde{p}_i$ a zero value of x_i, hence of (A_i, B_i), is obtained. Thus for all $p_i \geq \tilde{p}_i$ $(A_i, B_i) = 0$, while for any $p_i < \tilde{p}_i$ the value of (A_i, B_i) is unique and positive.

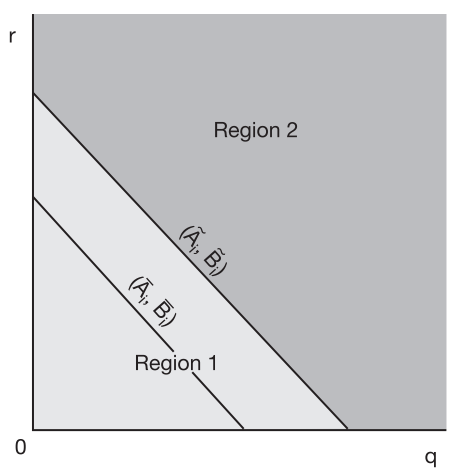

Figure 2.1 Individual demand function.

Let $(\tilde{A}_i, \tilde{B}_i)$ be the value of (A_i, B_i) corresponding to $\tilde{p}_i$. Then for $p_i = \tilde{p}_i$ (12) divides the first quadrant into two regions, as shown in Figure 2.1. Region 1 consists of all points (r, q) below this line and region 2 of all points on or above the line. Given a value of (r, q) in region 1, the corresponding value of (A_i, B_i) is uniquely determined by the line (12) which passes through the given point and (A_i, B_i) is positive. Given any value of (r, q) in region 2, the corresponding (A_i, B_i) is zero. Thus the individual demand function (6) is portrayed in Figure 2.1.

Any (A_i, B_i) in the range of (6) satisfies $B_i/A_i = T_i$. Figure 2.2 represents the range of (6) as a segment $0I$, of the straight line passing through the origin with slope T_i. Any given point, say $(\bar{A}_i, \bar{B}_i)$, on $0I$ may be labelled with the set of (r, q) which satisfies

$$(r, q) = f_{3i}^{-1} \ (A_i, B_i) \tag{13}$$

and this set is $\{(\bar{r}, \bar{q})\}$. The two limiting points are noteworthy. For $(A_i, B_i) = (0, 0)$, $\{(r, q)\}$ = region 2. For the maximum value of (A_i, B_i), namely, the point I, $\{(r, q)\} = (0, 0)$. This is the only point for which $\{(r, q)\}$ is unary.

Consider any point, say $(\hat{A}_i, \hat{B}_i)$, not in the range of (6), that is, not on $0I$. Then $\{(r, q)\}$ = the empty set. Therefore $(\hat{A}_i, \hat{B}_i)$ is not a feasible point, a fact rooted either in the consumer's trade-off among the attributes (if $(\hat{A}_i, \hat{B}_i)$ is not on the straight line passing through the origin with slope T_i) or in the fact that demand is satiable (if $(\hat{A}_i, \hat{B}_i)$ is on that line).

Thus, Figure 2.2 portrays the inverse of the individual demand function (13). The domain of (13) is $0I$ and there is only one point, namely I, for which (13) is single valued.

A given point, say $(\bar{A}_i, \bar{B}_i)$, in the domain of (13) corresponds to a set

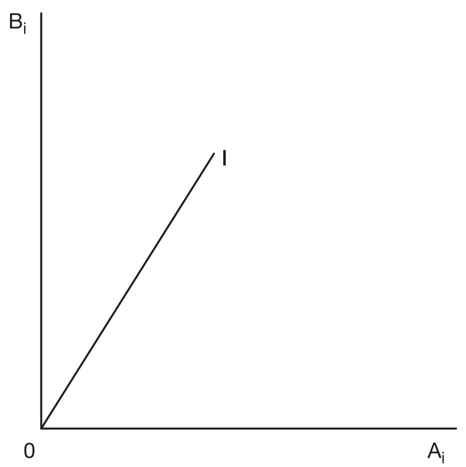

Figure 2.2 Inverse of individual demand function.

$\{(\bar{r}, \bar{q})\}$, but each element in that set satisfies (11), where $\bar{p}_i$ is unique. Moreover, $(\bar{A}_i, \bar{B}_i)$ corresponds to a unique value of x_i, say $\bar{x}_i$, given by

$$\bar{x}_i = \bar{A}_i / a_i = \bar{B}_i / b_i$$

Therefore, $\bar{R}_i = \bar{p}_i \bar{x}_i$, the consumer's total expenditure on the product, is also unique. Thus, it has been demonstrated that *the individual consumer's expenditure function*

$$R_i = f_{4i}(A_i, B_i) \tag{14}$$

is single valued.

5. Graphical representation of the market demand function

We now move directly from the case of one consumer to the most general situation: n consumers. In this section we derive a graphical representation of the market demand function (8). This is done for the purpose of delineating the multi-valued properties of the inverse market demand function (9) in Section 6.

Consider Figure 2.1 again. Such a portrayal of the individual demand function may be performed for every consumer, and the representations may take place on a single diagram, as in Figure 2.3. For each consumer (a typical

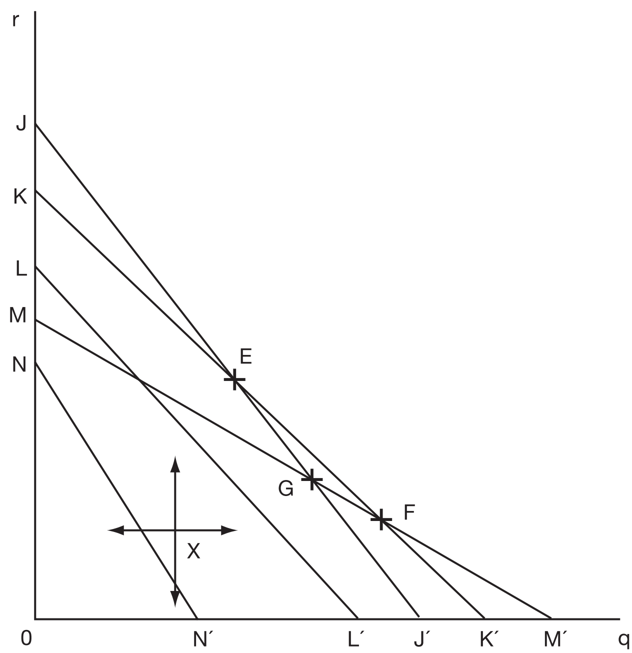

Figure 2.3 Market demand function.

one denoted as i), Figure 2.3 is everywhere dense with graphs of Equation (12). Since $(A_i, B_i) = 0$ for any point in region 2, (12) need be plotted and appropriately labelled (with the corresponding value of (A_i, B_i)) only for region 1 and the boundary between the two regions. Only one line (12) is actually drawn for a given consumer, the line (part of region 2) constituting the boundary between regions 1 and 2. This line is JJ' for consumer j, KK' for consumer k, etc.

Analogous to the case of one consumer, considered in Section 4 above, two regions can be defined with the joined line-segments $JEFM'$ as the boundary between them. For any (r, q) on or 'above' $JEFM'$, which is a short-hand way of saying 'for any (r, q) on $JEFM'$ or to the northeast of *any* point on $JEFM'$' or 'for any (r, q) *not* in the polygon $OJEFM'$ exclusive. of the outer part, $JEFM'$, of its perimeter,' (A, B) is zero. This is because (A_i, B_i) equals zero for every i. The complement of the closed set $OJEFM'$, together with the boundary $JEFM'$, is denoted as region II. Thus (A, B) is zero in region II.

The set $OJEFM'$ exclusive of the boundary $JEFM'$ is called region I. Any (r, q), say $(\bar{r}, \bar{q})$, in region I corresponds to a unique positive value of (A, B), say $(\bar{A}, \bar{B})$. The value is unique because the point $(\bar{r}, \bar{q})$ is on at most one line (12) for each consumer, and $(\bar{A}, \bar{B})$ is determined as $(\Sigma\, \bar{A}_i, \Sigma\, \bar{B}_i)$, where $(\bar{A}_i, \bar{B}_i)$ is the corresponding value of (A_i, B_i), namely, that value associated with Equation (12) which is satisfied by $(\bar{r}, \bar{q})$ or, geometrically, with the line (12) on which $(\bar{r}, \bar{q})$ is a point. As noted in Section 3, $(\bar{A}_i, \bar{B}_i)$ is unique for a given consumer i. The value of $(\bar{A}, \bar{B})$ corresponding to $(\bar{r}, \bar{q})$ in region I is positive because, by definition of region I, $(\bar{A}_i, \bar{B}_i)$ is positive for at least one i. If for a particular consumer i there is no line (12) on which $(\bar{r}, \bar{q})$ is a point, then $(\bar{A}_i, \bar{B}_i) = 0$. This assumes the convention of not considering (12) for a given consumer i as drawn in region 2 for that consumer. On the other hand, suppose (12) *were* plotted in region 2. Then if the line (12) satisfied by $(\bar{r}, \bar{q})$ is in region 2, that line is labelled as $(A_i, B_i) = (0, 0)$.

Thus every point (r, q) in the first quadrant of Figure 2.3 is labelled with the unique value of (A, B) corresponding to it. Hence, Figure 2.3 is a complete diagrammatic representation of the market demand function (8). Furthermore, it has been demonstrated anew that this function is single valued.

6. Multi-valued properties of the inverse of the market demand function

We wish to examine the single- cum multi-valued properties of (9), the *inverse* of the market demand function (8). This, too, will be done with the aid of Figure 2.3.

There are three regions to be considered, namely:

I_s = the set of (r, q) for which there is positive demand on the part of two or more consumers

I_m = the set of (r, q) for which there is positive demand on the part of one and only one consumer

II_m = the set of (r, q) for which there is positive demand on the part of no consumer

Referring to Figure 2.3, the polygon *OKEGFK′*, exclusive of the *KEGFK′* part of its perimeter, constitutes I_s; the triangles *JKE*, *EGF* and *FK′*M′, including *KEGFK′* but excluding the outer part of the perimeter of the polygon *OJEFM′*, namely, *JEFM′*, constitutes I_m, and region II constitutes II_m. To examine the properties of (9), we compare the label (i.e. value of (A, B)) attached to a given point in each of the regions I_s, I_m, II_m in turn, with that attached to any other point over all three regions.

6.1 The region I_s

Consider a given point $X = (r_X, q_x)$ in I_s. To this point corresponds a unique value of (A, B), say (A_X, B_X). We wish to compare (A_X, B_X) with the value of (A, B), say (A_W, B_W), associated with any other point, say $W = (r_W, q_W)$ in the domain of (r, q), i.e. in the first quadrant. Clearly, the latter point is on one (and only one) vector (with a given direction) protruding from X as the origin. Define all such vectors to be exclusive of the origin (the point X itself). There are four vectors of importance in delineating cases, namely, those due north, due south, due east and due west.

Case (i) w on the due west, due south, or any intermediate vector

(A_W, B_W) is unambiguously greater than (A_X, B_X). The proof is as follows. If W is on the due west vector, $q_W < q_X$ while $r_W = r_X$; if on the due south vector, $r_W < r_X$ while $q_W = q_X$; if on an intermediate vector, $r_W < r_X$ and $q_W < q_X$. In all three situations p_{w_i} is unambiguously less than P_{X_i} for every consumer i. Therefore $(A_{W_i}, B_{W_i}) \geq (A_{X_i}, B_{X_i})$ for every i. However, since $(A_{X_i}, B_{X_i}) > 0$ for at least one i (in fact, for at least two i), it must be that for these i $(A_{W_i}, B_{\mathrm{W}_i}) > (A_{X_i}, B_{X_i})$. It follows that $(A_W, B_W) > (A_X, B_X)$.

If X is on both the r and q axes, i.e. at the origin, then none of the vectors considered is defined. Such situations do not affect the demonstration, since, irrespective of the location of X, any other point in the domain of (r, q) is on a defined vector.

Case (ii) W on the due north, due east, or any intermediate vector

An argument precisely analogous to that used for case (i) demonstrates that $(A_W, B_W) < (A_X, B_X)$. This holds irrespective of whether W is in I_s, I_m or II_m. The important facts in the demonstration are that $p_{W_i} > p_{X_i}$ for all i and $(A_{X_i}, B_{X_i}) > 0$ for at least one i.

Case (iii) W northwest or southeast of X

The only remaining situations to consider are W in any vector intermediate between due north and due west or due south and due east. Associated with any such vector, considered as a straight line in the first quadrant, is its slope, the absolute value of which is denoted as T_{W^*}, the subscript W indicating correspondence to that point. A value of T_{W^*} corresponds to two vectors, one northwest, the other southeast of X, where the designations 'northwest' and 'southeast' exclude the due north, due west and due south, due east vectors, respectively. The following statements hold for all i. If W is northwest of X, $T_i > T_{W^*} \Rightarrow p_{W_i} < p_{X_i}$ and $T_i < T_{W^*} \Rightarrow p_{W_i} > p_{X_i}$. If W is southeast of X, $T_i > T_{W^*} \Rightarrow P_{W_i} > p_{X_i}$ and $T_i < T_{W^*} \Rightarrow p_{W_i} < p_{X_i}$. In either case $T_i = T_{W^*} \Rightarrow p_{W_i} = p_{X_i}$. We recall that

$$p_{W_i} < p_{X_i} \Rightarrow (A_{W_i}, B_{W_i}) \geq (A_{X_i}, B_{X_i})$$

$$p_{W_i} > p_{X_i} \Rightarrow (A_{W_i}, B_{W_i}) \leq (A_{X_i}, B_{X_i})$$

$$p_{W_i} = p_{X_i} \Rightarrow (A_{W_i}, B_{W_i}) = (A_{X_i}, B_{X_i})$$

In Case (iii) $(A_W, B_W) \neq (A_X, B_X)$. This will be proved by contradiction. Let W be northwest of X. If $(A_W, B_W) = (A_X, B_X)$ then, in particular, $A_W = A_X$ and $B_W = B_X$. Assume $A_W = A_X$. It will be shown that this implies $B_W \neq B_X$.

For every consumer i $B_i = T_i A_i$, irrespective of the value of A_i. Define $\Delta A = A_W - A_X$, $\Delta B = B_W - B_X$, $\Delta A_i = A_{W_i} - A_{X_i}$, and $\Delta B_i = B_{W_i} - B_{X_i}$. For ease in notation let p_* range over the set of i for which $\Delta A_i > 0$ and n_* range over the set of i for which $\Delta A_i < 0$. Now $\Delta A = \sum_{p_*} \Delta A_i + \sum_{n_*} \Delta A_i = 0$ by hypothesis. We want to prove that $\Delta B = \sum_{p_*} \Delta B_i + \sum_{n_*} \Delta B_i = \sum_{p_*} T_i \Delta A_i + \sum_{n_*} T_i \Delta A_i \neq 0$, i.e. that $\sum_{p_*} T_i \Delta A_i \neq -\sum_{n_*} T_i \Delta A_i$. For any i in the subset ranged over by p_*, $T_i > T_{W^*}$; for any i in the subset ranged over by n_*, $T_i < T_{W^*}$. By transitivity T_i for any i in the set ranged over by p_* exceeds T_i for any i in the set ranged over by n_*. By hypothesis $\sum_{p_*} \Delta A_i = -\sum_{n_*} \Delta A_i$. Therefore $\sum_{p_*} T_i \Delta A_i > -\sum_{n_*} T_i \Delta A_i$, providing only that both the set of i ranged over by p_* and the set of i ranged over by n_* are not empty. If both sets are empty, then $\sum_{p_*} T_i \Delta A_i = -\sum_{n_*} T_i \Delta A_i = 0$. However, the fact that X is in I_s eliminates the possibility that both sets are empty. There is positive demand at X on the part of at least two consumers. It is true that $T_i = T_{W^*} \Rightarrow \Delta A_i = 0$. However, by assumption $i \neq j \Rightarrow T_i \neq T_j$; therefore there cannot be more than one consumer for whom $T_i = T_{W^*}$. Hence, for at least one consumer i, $\Delta A_i \neq 0$. This fact combined with the hypothesis $\Delta A = 0$ implies that $\sum_{p_*} \Delta A_i, \sum_{n_*} \Delta A_i \neq 0$, that is, neither of the subsets of i can be empty. Thus it has been demonstrated that $A_W = A_X \Rightarrow B_W > B_X$ and, in particular, $B_W \neq B_X$. An analogous proof applies for W southeast of X.

In summary, for any given point (r, q) in I_s the corresponding value of (A, B) differs from that of any other point in the domain of (r, q).

6.2 *The region* $\mathbf{I}_m$

Consider a given point $Y = (r_Y, q_Y)$ in I_m. The unique value of (A, B), say (A_Y, B_Y), corresponding to this point is to be compared with that corresponding to any other point, say W, in the domain of (r, q). The latter point is on one (and only one) vector having Y as its origin. If W is on either (i), the due west, due south, or any intermediate vector or, (ii), the due north, due east, or any intermediate vector, then the proofs and results associated with X in I_s are directly applicable, with X replaced by Y. After all, these demonstrations were dependent only on $(A_{X_i}, B_{X_i}) > 0$ for at least one consumer i, a condition fulfilled in I_m.

Let k be the consumer the demand of whom is positive at Y, that is, $(A_{Y_k}, B_{Y_k}) = (A_Y, B_Y)$. If W is northwest or southeast of Y, two cases must be examined: $T_{W^*} \neq T_k$ and $T_{W^*} = T_k$.

Case (i) $T_{W^*} \neq T_k$

In this case $(A_W, B_W) \neq (A_Y, B_Y)$. The proof is similar to that used for X in I_s and that demonstration will be followed. Let W be northwest of Y. Assume $A_W = A_Y$. It will be shown that this implies $B_W \neq B_Y$. In the definition of the Δ variables replace X by Y. It need only be checked that both the set of i ranged over by p_* and the set of i ranged over by n_* are not empty. Recall that $(A_{Y_i}, B_{Y_i}) > 0$ for $i = k$ and $= 0$ for $i \neq k$. $\Delta A_k = 0 \Rightarrow T_{W^*} = T_k$, but the latter relationship is excluded by hypothesis. Therefore $\Delta A_k \gtrless 0$. Then the hypothesis $\Delta A = 0$ implies that $\Delta A_i \lesseqgtr 0$ for at least one other i. Hence $\sum_{p_*} \Delta A_i$, $\sum_{n_*} \Delta A_i \neq 0$. Thus it has been proved that $B_W \neq B_Y$. A similar argument applies for W southeast of Y.

Case (ii) $T_{W*} = T_k$

If $T_{W*} = T_k$, then $\Delta A_k = 0$. Providing W is sufficiently close to Y, for all $i \neq k$ $(A_{W_i}, B_{W_i}) = (A_{Y_i}, B_{Y_i}) = (0, 0)$ (with the latter equation true by hypothesis), hence $\Delta A_i = 0$ and $(A_W, B_W) = (A_Y, B_Y)$. The meaning of 'sufficiently close' is that W lies not merely in I_m but in that part of I_m in which k is the consumer with positive demand ('the applicable subregion'), which is obviously that subregion in which Y is situated. Actually the designation of the appropriate subregion is superfluous under the hypothesis $T_{W*} = T_k$. This is because (A_k, B_k) is constant along the vector of slope T_k passing through W and has a positive value of (A_{Y_k}, B_{Y_k}). W can be neither in II_m nor in any part of I_m in which k is not the consumer with positive demand; for these locations imply $(A_{W_k}, B_{W_k}) = 0$. Thus it suffices to say that W lies in I_m, providing the condition $T_{W*} = T_k$ is understood. For example, considering Figure 2.3, assume $k = J$. Then the applicable subregion is the triangle JKE inclusive of the side KE but exclusive of the side JE. If W is outside the applicable subregion and $T_{W*} = T_k$, then W must be in I_s (as is apparent from Figure 2.3). Then $(A_W, B_W) > (A_Y, B_Y)$, since $(A_Y, B_Y) = (A_{Y_k}, B_{Y_k})$, $\Delta A_k = 0$ and $(A_{W_i}, B_{W_i}) > 0$ for at least one $i \neq k$. Of course, the statement has already been proved in examination of the region I_s. (Notationally, transform W into X and X into W.)

In summary, for any given point (r, q) in I_m the corresponding value of (A, B) differs from that of any other point in the domain of (r, q), with the exception of all points in the applicable subregion of I_m that are on the line (12) that passes through the given point and pertains to the consumer with positive demand at that point.

6.3 The region II_m

By definition any point in II_m corresponds to $(A_i, B_i) = 0$ for all i, hence to $(A, B) = 0$. On the other hand, since for any point in I_s or I_m the corresponding $(A_i, B_i) > 0$ for at least one i, the (A, B) $(= (0, 0))$ associated with any point in II_m differs from the (A, B) $(> (0, 0))$ associated with any point in I_s or I_m.

Summary of multi-valued properties of (9)

Define the following regions in the A–B plane:

I_s^{-1} = the set of (A, B) given by (8) restricted to I_s
I_m^{-1} = the set of (A, B) given by (8) restricted to I_m
II_m^{-1} = the set of (A, B) given by (8) restricted to II_m

Recalling that $(A, B) = (\Sigma A_i, \Sigma B_i)$, equivalent definitions are:

I_s^{-1} = the set of (A, B) for which at least two of the n components (A_i, B_i), $(i = 1, \ldots, n)$ are positive

I_m^{-1} = the set of (A, B) for which one of the n components (A_i, B_i) is positive

II_m^{-1} = the set of (A, B) for which none of the n components (A_i, B_i) is positive, i.e. $II_m^{-1} = (0, 0)$

I_s^{-1}, I_m^{-1}, and II_m^{-1} are disjoint regions that exhaust the range of (8); but the range of (8) is the domain of (9). Therefore the above analysis provides a complete description of the single-valued cum multi-valued properties of (9), summarised as follows.

(1) Given any (A, B) in I_s^{-1}, the corresponding value of (r, q) is unique. Thus (9) restricted to I_s^{-1} is single valued and (8) restricted to I_s is one-to-one.

(2) Given any (A, B) in I_m^{-1}, there is a corresponding *set* of (r, q) (that is, a set given by (9)) each element of which satisfies

$$\bar{p}_k = ra_k + qb_k$$

where k is the consumer with positive demand at the given value of (A, B) and $\bar{p}_k$ is a particular value of p_k.

(3) Corresponding to $(A, B) = (0, 0)$ (the sole value of (A, B) in II_m) is the set of (r, q) each element of which satisfies

$$p_i = ra_i + qb_i \quad (i = 1, \ldots, n)$$

where $p_i \geq \tilde{p}_i$ and $\tilde{p}_i$ is the value of P_i beyond which $(A_i, B_i) = (0, 0)$.

The total-expenditure function

Clearly, the multi-valued nature of (9) in regions I_m^{-1} and II_m^{-1} is analogous to that of (13) in regions 1 and 2 of its range, respectively. Then one would expect that *the total-expenditure function (10) is single valued*, which is true and proved as follows.

For a given value of (A, B) in I_s^{-1}, (9) yields a unique value of (r, q). Then the value of R corresponding to (A, B) is also unique, because $R = rA + qB$ and each variable on the right-hand side has a unique value.

Corresponding to a given value of (A, B), say $(\bar{A}, \bar{B})$, in I_m^{-1} is a unique value of p_k, say $\bar{p}_k$, where k is the consumer with positive demand at $(\bar{A}, \bar{B})$, that is, $(\bar{A}, \bar{B}) = (\bar{A}_k, \bar{B}_k) > 0$. The given point also determines unique values of p_i, say $\bar{p}_i$, for all $i \neq k$. However, by hypothesis, (1) yields $0 = \bar{x}_i = f_{1i}(\bar{p}_i)$, all $i \neq k$, and $x_i = 0 \leftrightarrow (A_i, B_i) = 0$. Therefore R_k is uniquely given by $\bar{p}_k\bar{x}_k$ or, equivalently, by

$$R = R_k = f_{4k}(\bar{A}_k, \bar{B}_k)$$

Corresponding to (A, B) in II_m^{-1}, that is, to $(A, B) = (0, 0)$, $R = 0$ and is therefore unique.

7. Graphical representation of the inverse of the market demand function

Just as Figure 2.3 is a portrayal of (8), so it is desired to achieve diagrammatic representation of the inverse function (9). In principle the task is as straightforward as the derivation of Figure 2.2 given in Figure 2.1. In fact, it logically begins with a representation of (13) for each i as in Figure 2.2 but on a unique diagram. The problem is to convert the A_i–B_i plane to the A–B plane. This is done as follows. For a given value of (r, q) the corresponding (unique) values of $(A_i\ B_i)$ $(i = 1, \ldots n)$ are found on the diagram. Apply vector addition to these values of (A_i, B_i); the result is (A, B), and this point is labelled by the given value of (r, q). This may be performed for all values of (r, q); the result is a geometric representation of (9). Of course, some values of (A_i, B_i) will be associated with multiple values of (r, q). A unique labelling is obtained by converting the diagram to a representation of (10). In that case the domain (A, B) remains unchanged; but each point is identified by the total expenditure R pertaining to it.

While the above analysis is correct, it is too general to answer certain questions of importance in examining market equilibrium. For example, can a sufficiently small neighbourhood be drawn around any given point in the domain of (9) such that no other point in the domain is within that neighbourhood? In other words, is the domain a set of points all disconnected? Or, to take another extreme, is the domain a convex set? The general problem is, clearly, to determine the form of the set of points constituting the domain. This requires a detailed performance of the procedure outlined above. We begin with the case of two consumers ($n = 2$), and then extend the analysis to an arbitrary number of consumers.

7.1 The case of two consumers

Even for $n = 2$ we consider two sub-cases. First, we assume that neither (A_j, B_j) nor (A_k, B_k) is zero everywhere in region I_m. Later we remove this assumption.

Case (i): neither (A_j, B_j) nor (A_k, B_k) zero everywhere in I_m

As in Figure 2.3, (12) is plotted for region 1 and the boundary between regions 1 and 2 for each consumer j and k. In Figure 2.4 representative lines (12) are drawn from the boundaries to the origin. It is assumed that $T_j > T_k$. Hence the steeper set of lines pertains to consumer j. As in Figure 2.2, (13) is represented for $i = j, k$ as OJ and OK, respectively, in Figure 2.5.

The maximum values of (A_j, B_j) and (A_k, B_k), namely, those for $(r, q) = (0, 0)$, are represented by the points J and K in Figure 2.5. Applying (vectoral) addition to these points, the maximum value of (A, B) is obtained, and denoted as M. By a 'value of (A, B)' is understood, of course, a point in the

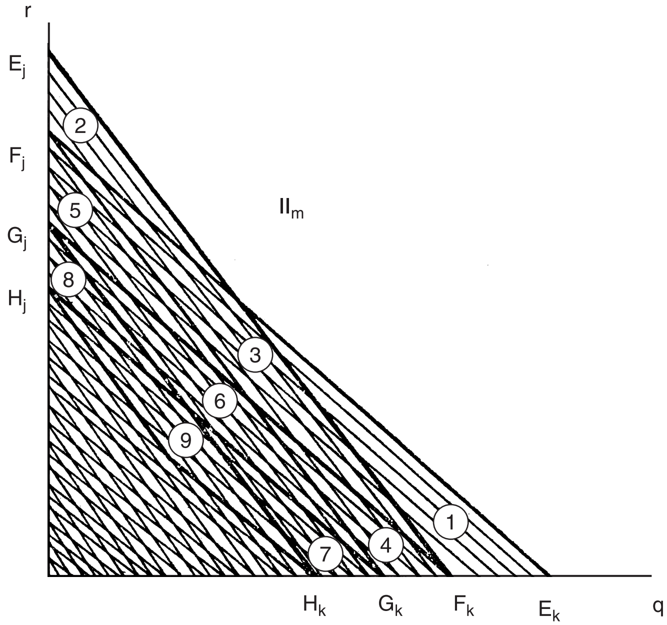

Figure 2.4 Market demand function.

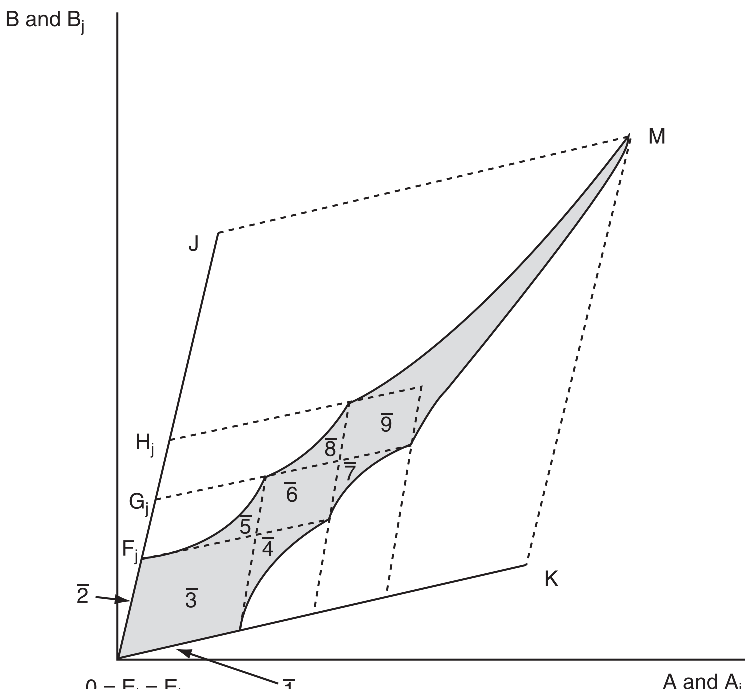

Figure 2.5 Inverse of market demand function.

domain of (9) or, equivalently, in the range of (8). Since every value of (A, B) is obtained by addition of points (A_j, B_j) and (A_k, B_k), and the latter points are on the lines $0J$ and $0K$, respectively, the domain of (9) is covered by the parallelogram $OJMK$. The latter object is the set of all sums $((A_j, B_j) + (A_k, B_k))$ such that (A_j, B_j) is a point on $0J$ and (A_k, B_k) a point on $0K$. Viewing Figure 2.4, it is clear that it is not true that every line (12) for consumer j intersects every line for consumer k. In fact, no line for either consumer intersects *every* line of the other consumer. Hence the domain of (9) is not the entire parallelogram $OJMK$ but only part of it. The maximum value of (A, B) is M; the minimum value is the origin (0) and applies for (r, q) in region 2 of both consumers. To discover the other points in the domain, a detailed analysis of Figure 2.5 is required.

The dark lines in Figure 2.4 separate I_s and I_m into subregions. These subregions are either triangles or parallelograms and are identified by a number. Also, II_m is demarcated by these lines. Each of the dark lines corresponds to a particular value of (A_j, B_j) or (A_k, B_k), as the case may be. These values are represented by appropriately subscripted capital letters which, for convenience, are printed along the vertical axis for consumer j and the horizontal one for consumer k. They are also indicated along the lines $0J$ and $0K$ in Figure 2.5. In this plane E_j and E_k coincide with the origin. The procedure is to map each region or subregion in the r–q plane (Figure 2.4) into the A–B plane (Figure 2.5). The corresponding regions in Figure 2.5 are indicated by the same numbers as in Figure 2.4, but with a bar above. Corresponding to 1 and 2 are the segments E_kF_k and E_jF_j, respectively, in Figure 2.5. Region 3 involves every line (12) corresponding to (A_j, B_j) between E_j and F_j, inclusive, intersecting every line corresponding to (A_k, B_k) between E_k and F_k, inclusive. Thus $\bar{3}$ is the set of all points (including the perimeter) in the parallelogram determined by 0, F_j and F_k. Region 4 consists of intersections of lines E_j to F_j with lines F_k to G_k. However, not all lines E_j to F_j intersect all lines F_k to G_k. F_j alone does that; E_j, on the other hand, intersects only F_k. There is a continuous increase in the lines (A_k, B_k) intersected as (A_j, B_j) moves from E_j to F_j. (Actually, the intersections of F_k with the E_j to F_j lines have already been counted in $\bar{3}$; but it clearly is irrelevant whether or not the parallelogram $\bar{3}$ is considered net of the outer half of its perimeter or, in general, to what regions boundary (dark) lines are assigned.) Addition of the appropriate points, thus determined, on $0J$ and $0K$ produces region $\bar{4}$. Region $\bar{5}$ is obtained analogously. Region $\bar{6}$ is the intersection of all lines F_j to G_j with all lines F_k to G_k. In Figure 2.5 this corresponds to the parallelogram with vertices $(F_j + F_k)$, $(G_j + F_k)$, $(G_j + G_k)$ and $(F_j + G_k)$. Regions $\bar{7}$, $\bar{8}$ and $\bar{9}$ are determined precisely analogously to regions $\bar{4}$, $\bar{5}$ and $\bar{6}$, respectively; in the above description the intervals F_j to G_j and F_k to G_k are replaced by G_j to H_j and G_k to H_k, respectively. The procedure of mapping triangles and parallelograms in the r–q plane into semi-triangles (that is, three-sided objects, at least two sides of which are straight lines) and parallelograms, respectively, in the A–B plane may be continued indefinitely. Point M will be reached as a limit. Thus the domain of (9) is the shaded area in Figure 2.5. Concerning the semi-triangles, a comment should be made about the side opposite the joint vertex of two parallelograms (called the 'third side'). In the diagram this side is drawn as a curved line convex to the vertex. Alternatively, the side may be a curve concave to the vertex or even a straight line. However, it must be downward-sloping with respect to that vertex (as already demonstrated), and the precise form is one of the three outlined and is identical for all semi-triangles. The latter circumstances are reflections of the fact that all triangular regions in the r–q plane are essentially of the same nature as regards the set of intersections of the j and k lines.

In summary, the domain of (9) is entirely within the parallelogram $OJMK$, includes the points 0 and M, and consists of a set of points that is closed and everywhere dense within (and on) its boundary. There are no discontinuities or gaps in the set constituting the domain. It is a connected set with minimum and maximum values of 0 and M, respectively. However, it is not necessarily a convex set.

Case (ii): (A_j, B_j) and (A_k, B_k) arbitrary in I_m

The assumption that both (A_j, B_j) and (A_k, B_k) have positive values in I_m may now be dropped. Suppose, for example, that (A_k, B_k) is zero everywhere in I_m. Then Figure 2.6 replaces Figure 2.4. The mapping from the r–q plane to the A–B plane may be performed in a manner similar to that outlined in detail above, that is, as a transformation of parallelograms and triangles into parallelograms and semi-triangles. Alternatively, the domain of (9) may be obtained by viewing it as the resultant of the intersection of successive individual j lines with successive *sets* of k lines. In the A–B plane (Figure 2.7) this corresponds to marking off points along each line parallel to $0K$ in the parallelogram $OJMK$. The only points on the lines from 0 to F_j constitute the segment $0F_j$ (= E_jF_j) itself. Then the relevant points on the parallel lines are intervals rather than single points until M is reached. The point set constituting the domain is the shaded area in Figure 2.7. It remains true that the domain is a connected set with minimum and maximum values of 0 and M, respectively. However, it includes no point (apart from 0) on the line $0K$.

7.2 The case of n consumers

The deductions regarding the shape of the mapping on the A–B plane remain valid for an arbitrary number of consumers. *The minimum value of the point set constituting the domain of the inverse of the market demand function is the origin* (*0*); the maximum value (*M*) *is calculated as the sum of all vectors 0I, I varying with i. A connected set joins the minimum and maximum points. This set is everywhere dense within (and on) its boundary. It contains no discontinuities or gaps*. The basic reason for these properties is that the individual demand functions (1), hence (6), are themselves continuous. In particular, they are continuous in the positive neighbourhood of $x_i = 0$ (or (A_i, B_i) = (0, 0)). It is

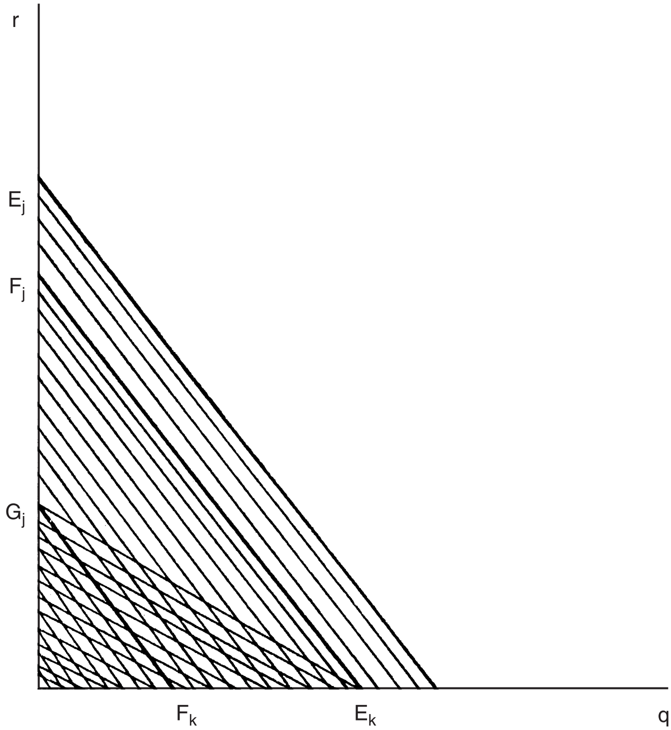

Figure 2.6 Market demand function.

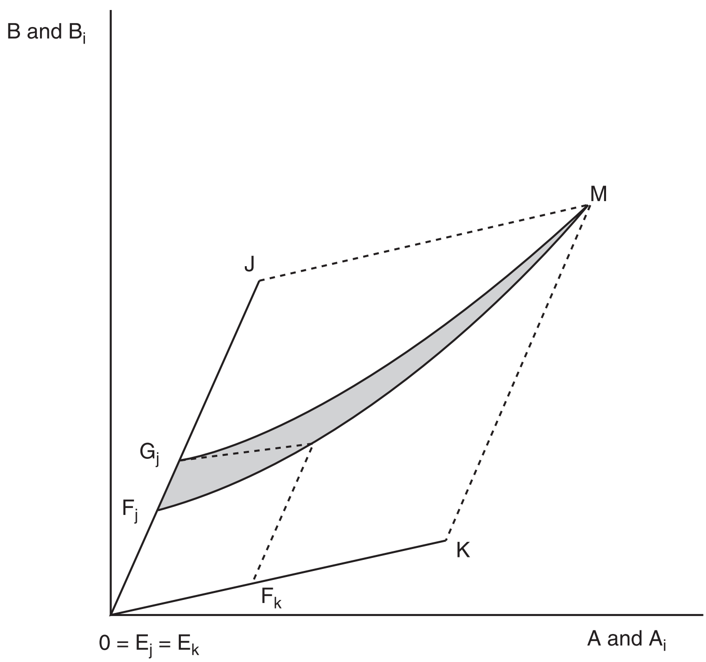

Figure 2.7 Inverse of market demand function.

this property which preserves the connectedness of the domain as (r, q) is altered to induce, on the part of a particular consumer, demand that is positive for the first time.

This completes the detailed description of the representation of the inverse demand function (9) in the A–B plane. Although the discussion has centred on the shape of the *domain* of (9), it should be recalled that every point on the domain is labelled by the corresponding set of values of (r, q). This set is not unary for all points, i.e. (9) is not single valued everywhere. However, if the labelling is changed to the corresponding value of R, the total-expenditure function (10) is obtained, and it is single valued.

8. Determination of equilibrium

Multi-dimensional pricing according to the attributes of a product results in a market demand function that is well-behaved and quite susceptible to analysis. The most important results of this paper are summarised in the statement that *the domain of the inverse of the market demand function is an everywhere-dense connected set with a unique total expenditure corresponding to any given point in the domain*. These properties give rise to the existence of equilibrium under the usual market structures.

The cost conditions for the suppliers of the product may involve individual costs per unit of production of each attribute (costs which may be zero) and joint (non-separable) costs of producing the attributes 'together'. If the attributes are produced in unvarying proportions, the 'supply curve' (i.e. the industry total-cost curve) is represented by a straight line through the origin with slope equal to the 'production trade-off' of the attributes. This line may be plotted in Figures 2.2, 2.5, or 2.7 above.

Then two cases may be distinguished. The 'supply curve' may or may not intersect the inverse of the market demand function. If this intersection occurs, the determination of the equilibrium output (A, B) and price (r, q) becomes a simple matter. Under pure competition, for example, the *maximum* (A, B) in the intersection constitutes the point of supply, unless such a point involves a negative profit for the marginal supplier. In the latter circumstances, the equilibrium point is the maximum among those points in the intersection for which this profit is non-negative. Under pure monopoly, in contrast, the supplier selects as the amount of his output that point (A, B) in the intersection which maximises his profit.

What happens in the case of a null intersection of the 'supply curve' and the inverse of the market demand function? Then one of the attributes is redundant, and under pure competition its price falls to zero. We recall that M is the point of maximum (A, B) in the domain of the inverse of the market demand function. If M yields a non-negative profit for the marginal supplier, it is the equilibrium point. If a negative profit is obtained, then supply is given by the point in the domain which involves maximum output, subject to non-negative profit for the marginal supplier. Under *either* equilibrium, suppliers obtain revenue from only the non-redundant attribute, its price given by the applicable element in the range of the inverse of the market demand function.

A null intersection under monopoly also gives a determinate solution. We are assuming that the monopolist throws onto the market the entirety of his production, resulting now in a zero price for the redundant attribute. Then he selects that point in the domain which maximises his profit, all of which emanates from the non-redundant attribute. The monopolist would be able to obtain a greater profit, however, by hoarding part of his output of the redundant attribute, so that he obtains a positive price for it. Yet he might refrain from doing so to avoid a loss in goodwill.

Thus far we have assumed that there is a fixed 'production trade-off' of the attributes. If this assumption is relaxed, the above analysis still holds. The 'supply curve' is represented no longer by a straight line but rather by a greater expanse. If the attributes can be produced in any combination at all, then the intersection of the 'supply curve' and the inverse of the market demand function becomes that inverse itself, and the analysis then proceeds as above.

Indivisibilities in production would give rise to further complications. Nevertheless, points of equilibrium would exist under both purely-competitive and monopolistic market structures.[3]

Notes

* Originally published in Willy Sellekaerts (ed.), *Econometrics and Economic Theory: Essays in Honour of Jan Tinbergen*, pp. 261–284, London: The Macmillan Press Ltd, 1974. Reproduced with permission of Palgrave Macmillan.

1 See Baumol [1], Lancaster [3], Simmonds [5] and the references cited therein.

2 'Constant coefficients' of attributes are also assumed by Baumol [1] and Lancaster [3].

3 For consideration of production indivisibilities in the case where the *domain* of the inverse of the market demand function can be treated as one-dimensional (i.e. $A = B$, in the terminology of the present paper), see Crew [2] and Officer [4].

References

[1] Baumol, William J., 'Calculation of Optimal Product and Retailer Characteristics: The Abstract Product Approach', *Journal of Political Economy*, LXXV (5) (October 1967) 674–85.

[2] Crew, Michael A., 'The Optimality of Pure Competition in the Capacity Problem: Further Comment', *Quarterly Journal of Economics*, LXXXIII (2) (May 1969) 341–3.

[3] Lancaster, Kelvin J., 'A New Approach to Consumer Theory', *Journal of Political Economy*, LXXIV (2) (April 1966) 132–57.

[4] Officer, Lawrence H., 'The Optimality of Pure Competition in the Capacity Problem', *Quarterly Journal of Economics*, LXXX (4) (November 1966) 647–51.

[5] Simmonds, Kenneth, 'Making Micro-Economics Multi-Dimensional', *Journal of Industrial Economics*, XVII (1) (November 1968) 18–28.

3 Monopoly and monopolistic competition in the international transportation industry*

There continues to be a dearth of attention paid to the transportation element in international-trade theory.[1] In this article I present a nonlinear-programming approach to the analysis of equilibrium in the international-transportation industry. However, the expression of transportation demand conditions in programming formulation (Section I and the Appendix), together with the presentation of cost conditions (Section II), would be applicable to various transportation situations, including both international and (with some qualifications) domestic trade.

The use of programming models in the analysis of commodity transportation between regions has appeared in earlier literature.[2] I explore a special case, in which equilibrium in commodity demand and transport is determined by the optimal price setting of a monopolistic transportation industry (Section III). While such a structure could exist in any international or domestic transportation industry, its epitome is the shipping conference *par excellence*, in which profit maximization is undertaken by a perfect cartel. As an application of the model, I examine the solution under an alternative market structure, a special form of monopolistic competition that might result from a breakdown of the monopoly (Section IV). Such a structure is presented as a more realistic alternative to the 'cut-throat competition' model that is generally viewed as the alternative to an effective conference system. The solutions under monopoly and monopolistic competition are compared (Section V), and it is shown that a collapse of the monopoly may indeed produce a market form of greater inefficiency, quite apart from problems of adjustment. Because of the direct applicability of the model to shipping conferences, transportation is considered oceanic, and the carrier of goods is denoted as (equivalently) a vessel, ship, carrier or liner. A summary of the results is presented in Section VI.

I. Demand conditions

The model is concerned with commodity trade between two countries, *A* and *B*. The rate of exchange between their currencies is fixed at unity. In order to delineate the fundamentals of the situation, I make the simplifying assumptions of constant costs of production (including any intracountry transportation charges), linear downward-sloping commodity demand functions, and pure competition in commodity markets. Income effects and interrelationships among commodity demand functions are ignored. With demand conditions assumed to be of the same nature for all commodities, the formal exposition in this Section can be confined to a single commodity. Notation for this commodity is as follows, with the definitions and subsequent relationships holding symmetrically for superscript *B*.

f^A = unit freight rate for transportation from *A* to *B*

p^A = unit cost of production in *A*

s^A = price in *A*

x^A = number of units imported by *A*

= number of units transported from *B* to *A*

y^A = number of units demanded in *A*

The domestic demand function in *A* is as follows:

$$y^A = \begin{cases} a^A + b^A s^A & \text{for } a^A + b^A s^A \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

with the parameters $a^A>0$, $b^A<0$.

The import demand function in *A* is:

$$x^A = \begin{cases} a^A + b^A (p^B + f^B) & \text{for } p^B + f^B \leq p^A \text{ and (simultaneously)} \\ & a^A + b^A(p^B + f^B) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The *B*-to-*A* transportation demand function is the above import demand function re-expressed so that the independent variable is f^B, the freight rate from *B* to *A*, rather than $(p^B + f^B)$, the landed price in *A*:

$$x^A = \begin{cases} (a^A + b^A p^B) + b^A f^B & \text{for } f^B \leq p^A - p^B \text{ and (simultaneously)} \\ & (a^A + b^A p^B) + b^A f^B \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Because the cost of production p^B is a parameter assumed to be unvarying (along with all other parameters of the model) whereas the freight rate f^B is a choice variable (as are all freight rates), it is only reasonable to study the import demand function in the form of the transportation demand function. The relationships among the domestic demand, import demand, and transportation demand function are illustrated in Figure 3.1.

The diagram consists of two partially superimposed sets of axes. First, consider the coordinate system with origin O. y^A is measured on the horizontal axis and s^A along the vertical axis. Plotting the function $y^A = a^A + b^A s^A$, its intercepts with the s^A and y^A axes are $-a^A/b^A$ and a^A, respectively. On the s^A axis the values of p^A and p^B are indicated, and horizontal lines extended to the right. p^B serves as the origin of the second coordinate system, with x^A and f^B measured on the horizontal and vertical axes, respectively.

With reference to the y^A–s^A axes, the following functions are depicted: the *domestic demand function* as the straight line from a^A to $-a^A/b^A$ joined with the part of the vertical axis extending from $-a^A/b^A$ to infinity, the *domestic supply function* as the horizontal line at p^A, and the *foreign supply function net of transportation cost* as the horizontal line at p^B. The *import demand function* is the straight line from a^A to F coupled with the vertical axis above p^A, in which case x^A is considered plotted along the y^A axis. The *transportation demand function* is the segment EF again coupled with the vertical axis above p^A, and this is with respect to the x^A–f^B coordinate system. Therefore the transportation demand function is the effective part of the import demand function, as imports cannot be obtained at a price lower than p^B.

The transportation demand function may be expressed as part of a formal programming system, in particular, as a set of constraints. For notational convenience, let

$$Q^B = f^B + p^B - p^A$$

$$R^A = a^A + b^A\,(p^B + f^B)$$

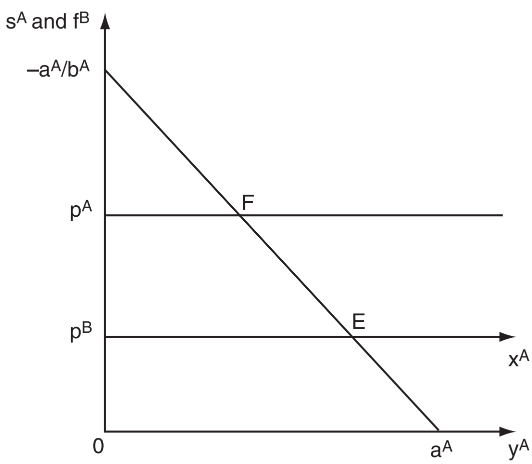

Figure 3.1 Demand functions.

Thus Q^B is the sum of the freight rate from B to A and the cost-of-production differential between A and B, while R^A is the number of units imported by A given that imports are positive (or, at the limit, zero). In this notation the demand function is:

$$x^A = \begin{cases} R^A \text{ for } Q^B \leq 0 \text{ and } R^A \geq 0 \\ 0 \quad \text{otherwise} \end{cases}$$

The set of constraints that constitutes the demand function follows:

$$x^A \geq 0 \tag{1}$$

$$x^A (x^A - R^A) = 0 \tag{2}$$

$$Q^B x^A \leq 0 \tag{3}$$

$$(z^B - Q^B)(z^B - L^B) = 0 \tag{4}$$

$$R^A (z^B - Q^B) \leq 0 \tag{5}$$

$$R^A (z^B - L^B) \leq 0 \tag{6}$$

$$R^A (x^A - R^A) \leq 0 \tag{7}$$

$$R^A z^B (x^A - R^A) \leq 0 \tag{8}$$

$$Q^B Q^B L^B + x^A - R^A \geq 0 \tag{9}$$

where L^B is a (any) predetermined number such that:

(i) $L^B > 0$

(ii) $L^B \geq Q^B$

(iii) $L^B \geq R^A/Q^B Q^B$ for $R^A, Q^B > 0$

and z^B is an instrumental variable created to aid in the construction of the function.

Relationships (1) through (9) specify the transportation demand function. A proof is provided in the Appendix.

The assumption of constant costs (horizontal supply functions) ensures that demand in A is met either entirely from home production (if $f^B > p^A - p^B$) or entirely from imports (if $f^B \leq p^A - p^B$).[3] Now consider the case of n commodities, with a typical commodity denoted as i. Then every superscripted symbol in relationships (1) through (9) may be given a subscript i to refer to commodity i. Furthermore, reversing the superscripts everywhere would switch the direction of the transportation demand function (A to B rather than B to A). However, the number of applicable demand functions is less than $2n$, in fact no greater than n. The reason is that a commodity i is potentially transportable from B to A only if $f_i^B \leq p_i^A - p_i^B$ and potentially transportable from A to B only if $f_i^A \leq p_i^B - p_i^A$, both of which conditions cannot hold simultaneously under any rational transport-pricing scheme ($f_i^A, f_i^B > 0$). Thus, given only the parameters p_i^A, p_i^B, the applicable one of the two superscripting schemes may be assigned to the relationships (1) through (9). This procedure may be followed for each commodity i, and the set of all constraints so chosen will be called *the applicable demand conditions*. It will be supposed that there are u commodities potentially transportable from B to A, a typical such commodity denoted as j, and v commodities potentially transportable from A to B, a typical one denoted as k. Then $u + v \leq n$.[4]

II. Cost conditions

Commodities have certain physical attributes which are of importance in oceanic transportation. These are the weight and volume dimensions. Thus let

w_i = number of weight tons per unit of commodity i

m_i = number of measurement tons[5] per unit of commodity i

A ship has well-defined limits to both weight and volume of cargo. For simplicity, I assume that only one type of vessel exists. Its capacity characteristics are:

W = maximum number of weight tons of cargo

M = maximum number of measurement tons of cargo

The selection of a given type of vessel is equivalent to choosing the production function itself (as distinct from the point of operation on it), as emphasized by Sturmey [12, pp. 15–16]. The optimal choice is not easily made, because of the diverse nature of the factors involved. ‘In general, the search for the lowest cost solution to a marine transport problem involves consideration of ship size, speed, equipment and manning. The *liner* owner, however, has in addition to consider the needs of the trades which he serves, any constraints imposed by the capacity of the ports in that trade and the conference rules’ [12, p. 5]. I assume that this efficiency problem has been solved. This is not to deny that the vessel type selected has an important effect on shippers’ welfare. For example, it has been suggested [15, p. 21] that under certain circumstances shippers (and presumably the ultimate consumers of their goods) might prefer to have service provided by slower vessels concommitant with lower freight charges.

There are three types of cost associated with the provision of oceanic transportation services: commodity costs, vessel expenses, and overhead expenses.[6] *Commodity costs*[7] are allocable to particular cargo. They consist predominantly of loading and unloading charges and special handling expenses, but include all costs that can be associated with specific freight. Commodity costs refer to a one-way voyage. They may differ for a particular commodity depending on the direction of trade (because loading and unloading expenses may differ in the two countries), but these costs per unit of a commodity (for a given direction – either A to B or B to A) are assumed independent of the amount of the commodity carried.

Vessel expenses[8] refer to the actual running of the vessel. They include such items as wages of the crew, fuel costs, and port dues. In general, all costs marginal to a voyage but not allocable to specific cargo are considered vessel expenses. While, clearly, these expenses vary as a function of the route of the vessel, it is reasonable to assume that they are constant for a given route for a given time period, irrespective of the quantity or quality of cargo. Since the model assumes a particular route, vessel expenses are constant.

Overhead expenses[9] are analogous to the conventional concept of fixed costs. They consist of expenses that occur (and are constant) irrespective of whether the vessel is actually operating on the route. Depreciation is the chief component of these charges.

Vessel and overhead expenses differ from commodity costs in two important respects. First, they cannot be allocated to specific commodities; they are joint rather than individual costs. Second, they apply to a round-trip rather than one-way voyage. Furthermore, it is known that vessel and overhead expenses are the preponderant part of total costs. Ferguson et al. state that 'The ratio of variable costs to fixed costs is small in shipping' [1, p. 323]. Quantitative measures of the importance of vessel and overhead expenses in total costs range from McLachlan's estimate of 65 percent [8, p. 328] to UNCTAD's figure of 75 percent [15, p. 7] and Sturmey's estimates of 75–80 percent [12, pp. 7–8] and 80 percent [11, p. 191].

The cost parameters may be expressed in symbols as follows:

c_j^B = commodity costs per unit of commodity j transported from B to A

c_k^A = commodity costs per unit of commodity k transported from A to B

V = vessel expenses of an A–B voyage

D = overhead expenses of an A–B voyage

The freight rates f_j^B and f_k^A are variables rather than parameters. However, one property of their equilibrium values can be presented a priori. It is inconceivable that any rational pricing system could entail a freight rate below commodity costs, as this would be equivalent to pricing below marginal costs.[10] Therefore the following relationships hold.

$$f_j^B = c_j^B + g_j^B \qquad j = 1, \ldots, u \tag{10}$$
$$f_k^A = c_k^A + g_k^A \qquad k = 1 \ldots, v$$

where $f_j^B, f_k^A > 0$, $g_j^B, g_k^A \geq 0$.

Equations (10) may be taken as definitions of g_j^B and g_k^A. These variables are that part of the freight rate above commodity costs, and so represent the commodity's contribution to the carrier's vessel and overhead expenses and profits (if any).

There remains an important variable to be defined:

N = number of vessels on the A–B route

x_j^A and x_k^B denote *total* transportation of commodities j and k, respectively (in their respective directions), that is, carriage by all vessels. W, M, V, and D, on the other hand, are properties of an individual vessel. In this connection the model incorporates the following assumptions:

1. Each vessel carries an identical amount of each type of cargo, namely, x_j^A/N, $j = 1, \ldots, u$, and x_k^B/N, $k = 1, \ldots, v$.
2. Vessel and overhead costs are additive. This holds even if ships are under the same management-ownership. Thus the vessel and overhead expenses of operating N ships are NV and ND, respectively.
3. The model incorporates a time period covered by a round-trip voyage performed by all ships. Demand conditions repeat themselves every time period.
4. The parameters listed below are taken as given, with their respective signs indicated.

$$a_j^A, c_j^B, p_j^A, p_j^B, w_j, m_j > 0, b_j^A < 0 \qquad j = 1, \ldots, u$$
$$a_k^B, c_k^A, p_k^A, p_k^B, w_k, m_k > 0, b_k^B < 0 \qquad k = 1, \ldots, v$$
$$W, M, V, D > 0$$

This leaves the variables f_j^B, g_j^B, x_j^A, $j = 1, \ldots, u$; f_k^A, g_k^A, x_k^B, $k = 1, \ldots, v$; and N as undetermined. Now, given the values of g_j^B and g_k^A, those of f_j^B and f_k^A are determined by equations (10). These values of f_j^B and f_k^A inserted into the applicable demand conditions determine the values of x_j^A and x_k^B. Thus, to complete the model, the only variables that need to be determined are the freight rates g_j^B, $j = 1, \ldots, u$ and g_k^A, $k = 1, \ldots, v$ and the number of vessels N.

Traditionally, the pricing problem in oceanic shipping has been viewed as a matter of selecting the choice variables g_j^B and g_k^A so as to cover vessel and overhead expenses. As Sturmey states: 'The problem in rate determination is to allocate overheads which are the major part of total cost. There is no "right" solution to this problem so that in an important sense all rates must be arbitrary' [12, p. 2]. While freight rates are arbitrary in the sense that any of a large number of rate structures could cover vessel and overhead costs, they are by no means indeterminate. To determine the freight rates g_j^B and g_k^A, one must superimpose a particular market structure on the model. In this respect, determination of the number of vessels N is an integral part of the rate-determination process. Equilibrium under two alternative market forms is examined in detail – monopoly in Section III and monopolistic competition in Section IV, with their results compared in Section V. The market forms are discussed in the context of shipping conferences.[11]

III. Equilibrium under monopoly

Monopoly in the provision of transportation services may refer to either of two kinds of market situations. It may denote classical monopoly, that is, a single ownership-management of all vessels on the route. Alternatively, it is a synonym for the conference system in its purest form – a perfect cartel. Indeed, as McLachlan points out, these are the two market forms immediately suggested by the nature of oceanic transportation: 'In a situation in which fixed costs predominate and total costs are not easily allocable among the various commodities moving in a given trade, stability of freight rates and services are unlikely to be provided except by the emergence of either a single firm monopoly or a collective monopoly' [8, p. 333]. The only difference between the single firm (classical monopoly) and the conference system *par excellence* (collective monopoly) is the retention of separate firm ownership and management of vessels under the latter structure. Then my assumption of additivity of vessel and ownership expenses ensures that the one situation may be identified with the other.

What are the functions of a conference? Ferguson et al. state its role succinctly. 'Directly or indirectly, conferences perform at least four major functions: (1) price-setting; (2) allocating output among the members; (3) dividing revenues; and (4) controlling entry' [2, p. 356]. Assume that the conference (or monopolist) operates its choice variables so as to maximize profits and that its control of entry is complete.[12] Then market equilibrium is determined as follows.

The conference selects $g_j^B, j = 1, \ldots, u$; $g_k^A, k = 1, \ldots, v$; and N to maximize

$$\pi_1 = \sum_j g_j^B x_j^A + \sum_k g_k^A x_k^B - N(V + D) \tag{11}$$

subject to the applicable demand conditions, (10), and also

$$\sum_j w_j x_j^A \leq NW, \sum_j m_j x_j^A \leq NM \tag{12}$$

$$\sum_k w_k x_k^B \leq NW, \sum_k m_k x_k^B \leq NM \tag{13}$$

$$N \geq 0 \tag{14}$$

Relationships (12) and (13) are the capacity constraints. The problem is equivalent to creating a 'giant' vessel N times the given size of an actual vessel. Strictly speaking, constraint (14) should be 'N a non-negative integer.' However, if one were to drop the assumption that vessel size is fixed, the problem could be treated as a choice of fleet rather than vessel capacity, and N could legitimately be taken as continuous.[13] The remainder of the article proceeds on this basis.

The objective function (11) may be rewritten as:

$$\pi_1 = N\pi_2 \tag{15}$$

where

$$\pi_2 = \sum_j g_j^B x_j^A / N + \sum_k g_k^A x_k^B / N - (V + D) \tag{16}$$

is not only the *average* profits per vessel but also the *actual* profits. This is because of the assumption that the cargo content of each vessel is the same, namely, x_j^A/N, $j = 1, \ldots, u$ and x_k^B/N, $k = 1, \ldots, v$, respectively. Thus the problem of output and revenue allocation among the conference member firms is automatically solved. Each member receives a share of the cargo in proportion to the number of its vessels on the route, and – with a unique ship type – each vessel receives the same share of cargo. Revenues (together with costs and profits) are then automatically distributed along the same proportions. Thus the conference is comparable to a multi-plant firm in which all plants are identical in cost conditions. There would be no incentive to move away from an equal allocation of output among the plants.[14]

Perhaps the most sophisticated model of shipping conferences extant is that of Abrahamsson [1]. Yet, in contrast to the above approach, his model is deficient in three important respects. First, Abrahamsson does not differentiate between commodities. He has only two commodities, namely, liner and tramp cargo, and the latter is obtained by the conference members (liners) entering the non-conference (tramp) market as price-takers. Thus only a single freight rate (comparable to f_i), that on liner cargo, is determined by the conference, allowing no scope for 'commodity discrimination' in price-setting. By commodity discrimination I mean having the freight rate g_i depend not only (or even necessarily) on the volume and weight capacity the commodity takes up but also on its particular demand conditions.[15] Second, Abrahamsson does not take account of the two-dimensional nature of ship capacity; both weight and volume capacities are subsumed under the category 'cargo space,' and hitting either constraint would make a vessel fully loaded. Third, Abrahamsson pays no attention to the spatial aspect of demand, to the fact that vessels operate on a given voyage consisting of several legs. The cargo on the *B*-to-*A* part of the round-trip voyage in my model does not compete for space with the cargo on the *A*-to-*B* part; yet each has the potential of contributing to the joint costs of the voyage, namely, vessel and overhead expenses. This is a case of complementary demands or what has been termed the peak-load pricing problem.[16] Nothing can be said a priori about the relative contribution of each direction of the voyage to vessel and overhead expenses. The share of either can range from zero to $V + D$ itself (and more if profits are earned).[17] Furthermore, the model does not imply that vessels will be fully loaded in either or both directions. Indeed, in Section V below I analyze a situation in which the parameters of the model are such that profit-maximization involves partially loaded vessels in both directions, with a capacity constraint potentially reached only as a limiting case.

The profit-maximization approach to the analysis of conference behavior has been rejected by Sturmey [11], who argues that the pricing policy of conferences is deterrent in nature. 'Pricing is, therefore, determined on a basis which will yield reasonable profits to the lines in the trade but will not attract extra competition' [11, p. 202]. It would be unreasonable to deny that deterrent pricing may be a *feature* of conference freight-rate determination. Indeed, observers writing earlier than Sturmey recognized that fact.[18] If the conference control of entry is complete, then deterrent pricing is inapplicable. If entry cannot be controlled fully and the conference members foresee the possibility of outside competition, then a deterrent pricing policy is not only applicable but also consistent with profit maximization. In the context of my programming model, one would simply add constraints of the form $f_i \leq h_i$, where h_i is the rate above which outside vessels would become effective competitors in the transportation of commodity *i*. If the conference has assessed such rates correctly (together, of course, with all other parameters of the model), then deterrent pricing in this form is an integral part of profit-maximization. Viewed in this light, deterrent pricing is analogous to the use of 'fighting ships,' with the latter technique having been outlawed by the United States since 1916.

Let us return to consideration of the maximization of (15). This may be envisaged as a two-stage procedure. The first stage is the construction of the function $\pi_{2x}(N)$, where π_{2x} is the maximum value of π_2 subject to the applicable demand conditions and constraints (10), (12) and (13). Though (12) and (13) as written refer to aggregates, they may be restated as pertaining to individual vessels by division by N, transforming x_j^A, x_k^B into x_j^A/N, x_k^B/N, respectively, and aggregate capacity limits into corresponding individual vessel limits. The second stage of the procedure is the selection of N according to the objective of maximizing $N\pi_{2x}$.

Analysis of the monopolistic-competition market form in Section IV requires knowledge of the form of the function $\pi_{2x}(N)$. Consider the demand functions for transportation services faced by individual vessels as themselves functions of the number of ships on the route. These demand functions are obtained from corresponding aggregate demand functions by transforming x_j^A and x_k^B into x_j^A/N and x_k^B/N, respectively. This means that the applicable demand conditions specify the demand functions faced by an individual vessel, providing the underlying functions $y_j^A = a_j^A + b_j^A s_j^A$ and $y_k^B = a_k^B + b_k^B s_k^B$ are replaced by $y_j^A = (a_j^A + b_j^A s_j^A)/N$ and $y_k^B = (a_k^B + b_k^B s_k^B)/N$, respectively. Thus the underlying functions rotate about the point $(s_j^A, y_j^A) = (-a_j^A/b_j^A, 0)$ or $(s_k^B, y_k^B) = (-a_k^B/b_k^B, 0)$ as N changes. The functions move inward (toward the origin) as N increases. These revised demand conditions will be called 'the applicable demand conditions for an individual vessel.'

To examine the implications of this fact on the π_{2x} function, a diagrammatic illustration is useful. An arbitrary commodity (nonsubscripted) and a given direction (B to A, as in Section I) are considered, but the principles gleaned again apply to all commodities and both directions. Figure 3.2 portrays the transportation demand function faced by an individual vessel under two alternative values of N. The segment EF and the vertical axis above p^A is the function for $N = N_c$; the segment GH and the same part of the vertical axis is the function for $N = (N_c - 1)$.

Drop perpendiculars from F and H to the x^A axis and denote their intercepts with the axis as F' and H', respectively. By an *attainable value* of x^A is meant a value in the positive part of the range of the transportation demand

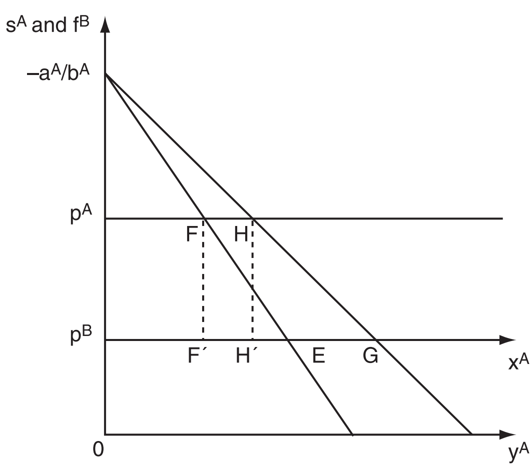

Figure 3.2 Transportation demand function.

function. The set of attainable values under N_c and $(N_c - 1)$ are the closed intervals $F'E$ and $H'G$, respectively. The union interval $F'G$ may be divided into three sub-intervals: $F'H'$, $H'E$, and EG, corresponding to points that are attainable under N_c but not $(N_c - 1)$, under both N_c and $(N_c - 1)$, and under $(N_c - 1)$ but not N_c, respectively. For the common sub-interval $H'E$ the demand price under $(N_c - 1)$ is greater than that under N_c. Therefore the sole decrement in the profit possibilities of the vessel is the loss of $F'H'$ as part of the attainable interval of x^A. This phenomenon occurs only because the model does not permit rationing of demand: for a given rate only the amount given by the demand function – and no less than that amount – is tenable. It is noted that the interval $F'H'$ decreases as p^A increases, and disappears for $p^A \geq -a^A/b^A$.

There is a very strong presumption that π_{2x} is a declining function of N. As N decreases, the demand for transportation of any commodity faced by each vessel unambiguously increases (or, at the limit, remains constant at zero) for every freight rate. Consideration of the inverse demand function reveals that the minimum positive value of demand will increase in certain cases. (In Figure 3.2 it increases from F' to H'.) This feature can be of relevance only if a capacity constraint is binding; but even here, in view of the general outward shift of all demand functions, appropriate reshuffling of rates, decreasing and perhaps eliminating transportation of some commodities, is likely to increase profits. Thus perceiving that a decrease in N creates economies of concentration on particular commodities in the provision of transportation services, this likelihood becomes a virtual certainty.

Therefore it is only reasonable to conclude that

$$\Delta \pi_{2x}/\Delta N < 0 \text{ for } N \geq 0$$

is a property of the π_{2x} function. Thus π_{2x} is a monotonically-decreasing function of N, a result incorporated in the next Section.

IV. Equilibrium under monopolistic competition

Suppose that a breakdown of the transportation-industry monopoly were to occur. What alternative market structure might take its place? The traditional view is that cut-throat or ruinous competition would result. The transportation demand functions for an individual vessel are downward-sloping, as shown in the previous section, only insofar as the owner of the vessel is faithful to the conference agreement. If he succeeds in cutting freight rates below that established by the conference and is the only conference member to do so, the demand functions for his ships become infinitely elastic. As McLachlan notes: '. . . while a liner conference as a whole faces an extremely inelastic demand curve, it is important to remember that the demand curve facing the individual firm within a conference becomes extremely elastic once that firm decides to cut rates secretly and provided that other members of the conference are not doing likewise' [8, pp.332–33]. If such behavior becomes known to other members, the conference itself may collapse to be replaced by a rate war among the former members of the conference. With low marginal (commodity) costs, cut-throat competition might occur in the sense that freight rates fall to such levels that vessel and overhead expenses are far from fully covered. In the short run ships may ply the route as long as their revenues are sufficient to pay for any part of their fixed costs (vessel and overhead expenses). In the long run such a situation cannot endure; firms will pull their vessels out of the route until the number of vessels is small enough so that all earn positive profits. Then the rate war might begin anew. Ultimately the risk of loss in such an unstable situation might be viewed by liner owners as too great to warrant any servicing of the route at all. Such an analysis has been the traditional justification for the conference system.[19] However, both the realism of the cut-throat competition model and its use as a justification for a conference monopoly have been discredited by Ferguson et al. [2, pp. 312–39 and 406–13] and Sturmey, in a change of view [11, pp. 191–96].

If cut-throat competition is not a likely alternative to a breakdown of the conference, what other market form might take its place? I suggest that the conference in form and substance would remain, except that its control of entry fails, to be replaced by unlimited free entry. In other words, the number of ships (N) is no longer under the control of the conference. Competitors simply are admitted into the conference. Thus suppliers of transportation services continue to collude in the setting of freight rates, but they cannot prevent free entry into the industry. The manager of each vessel selects g_j^B, $j = 1, \ldots, u$; g_k^A, $k = 1, \ldots, v$, to maximize π_2 subject to 'the applicable demand conditions for an individual vessel,' (10), (12), and (13), the latter two constraints transformed by division of N, as outlined in Section III. I denote this special market form as monopolistic competition.

Opinions might differ as to the realism of such a model, i.e., as to the openness of the conference to new members. Sturmey has a pessimistic view on the effectiveness of outside firms in thwarting the entry barriers: 'New lines can only be admitted to the conference on a vote of existing members. The meetings of the conferences are private and the conference decisions are usually stated either without explanation or with a stereotyped explanation without details' [10, p. 325]. Ferguson et al. are more sanguine about the ability of new firms to join the conference: 'It is now probably easier for a substantial liner company to enter a conference than it once was; but the analysis of conferences suggests that their entry would be difficult to prevent in any case' [2, p. 404]. However, Ferguson et al. do not analyze the consequences of such entry. I turn to a consideration of this matter in the light of my programming model.

The number of vessels, N, now enters the problem as a parameter, not a choice variable. This means that rules of behavior are necessary in order to determine the entry and exit of vessels and obtain equilibrium in the market. Since π_2 is envisaged as maximized given N, use may be made of the function $\pi_{2x}(N)$. Then the rules regarding change in N are as follows:

$\pi_{2x}(N) > 0 \Rightarrow N$ increases

$\pi_{2x}(N) < 0 \Rightarrow N$ decreases

Entry takes place if there are profit opportunities and exit if there are losses. The resulting market equilibrium is illustrated in Figure 3.3. The rules of behavior lead to N_2 as the equilibrium value of N, yielding zero profits per vessel. With a fixed vessel size, the solution would be the integral value of N less than, but closest to, N_2, involving positive profits in equilibrium.

It should be explained why this market form is called monopolistic competition. Its suppliers are colluding price-makers with mutually-dependent demand curves (in the sense that changes in the number of suppliers (vessels on the route) alter the demand curves faced by each vessel). The services sold are not differentiated but are identical, only because the demand (and cost) conditions are the same for every supplier.

Finally, it is noteworthy that Sturmey suggests cut-throat competition not as an alternative to monopolistic competition but rather as a *result* of open admission: 'If the conference admits new members freely it is likely to promote a rate war within the conference due to over-tonnaging . . .' [10, p. 343]. This assertion is in contradiction to my analysis, which shows that monopolistic competition results in a stable equilibrium.

V. Comparison of monopoly and monopolistic competition

In this Section, I compare equilibria under the two market forms developed above, namely, monopoly and monopolistic competition. Formally the

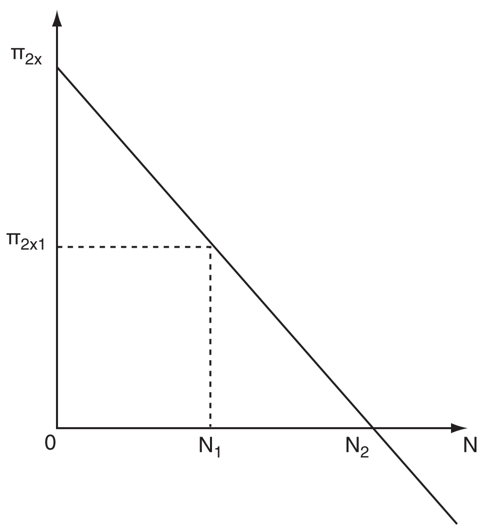

Figure 3.3 Per-vessel profits related to number of vessels.

problems differ in two respects: the objective function and the set of choice variables. However, since N is a parameter under monopolistic competition and because π_1, given that fact, is a multiple of π_2, π_2 may be replaced by π_1 as the objective function under monopolistic competition. Therefore there is one substantive difference between the two market structures: N is a choice variable under monopoly but a parameter under monopolistic competition, a parameter the equilibrium value of which is determined by certain rules of behavior.

Viewing the matter in a systematic fashion, monopoly involves maximization of $N\pi_2$, while monopolistic competition entails minimization of π_2 subject to the constraints (i) $\pi_2 \geq 0$ and (ii) $N \geq 0$. Underlying both problems, of course, is the set of common constraints, namely, the applicable demand conditions, (10), (12) and (13).

Under monopolistic competition, π_2, given N, is maximized with respect to g_j^B, $j = 1, \ldots, u$ and g_k^A, $k = 1, \ldots, v$, but π_2 is minimized with respect to N (subject to constraints (i) and (ii)). This is envisaged as a two-stage procedure. The first stage is the construction of the function π_{2x}; the second is the determination of a point on this function. Monopoly may be viewed as entailing the same two-stage procedure. It is the criterion for selection in the second stage that differentiates equilibrium under the two market forms. Given monopolistic competition, the equilibrium values of π_{2x} and N are determined by free entry, which changes N until π_{2x} is minimized subject to constraints (i) and (ii). On the other hand, the conference (or monopolist) selects N according to the objective of maximizing $N\pi_{2x}$. Formally this is performed without constraints. Clearly, however, unconstrained maximization of $N\pi_{2x}$ entails the satisfaction of (i) and (ii). Thus these constraints may be imposed upon the conference with no restrictive effect on behavior. Therefore the one difference between monopoly and monopolistic competition is that the former maximizes $N\pi_{2x}$ while the latter minimizes π_{2x}, both subject to the same set of constraints.

In view of this one difference, Figure 3.3 may be used to illustrate monopoly equilibrium, thus effecting a graphical comparison with monopolistic competition. The maximum value of $N\pi_{2x}$ is achieved by selecting the point of unitary elasticity on the π_{2x} function. N_1 is the equilibrium value of N and $N_1\pi_{2x1}$ the total profits. In the case of a fixed vessel size, the equilibrium (maximum) value of $N\pi_{2x}$ would be determined by selecting the rectangle with largest area among those with a diagonal from the origin to a point on the function.

Letting m represent 'equilibrium value under monopoly' and mc 'equilibrium value under monopolistic competition,' it may be generalized that:

$$(N)_m \leq (N)_{mc}$$

$$(\pi_2)_m \geq (\pi_2)_{mc}$$

$$(N\pi_2)_m \geq (N\pi_2)_{mc}$$

A further comparison between monopoly and monopolistic competition which is of interest is the equilibrium freight rates themselves: $g = (g_1^B, \ldots, g_u^B, g_1^A, \ldots, g_v^A)$. Replacing the dependent variable (π_{2x}) in the $\pi_{2x}(N)$ function by the corresponding g creates a function $g_x = g_x(N)$ defined as follows: for a given value of N, say N_k, g_x is the value of g that maximizes π_2, thus yielding $\pi_{2x}(N_k)$.

Construction of the $g_x(N)$ function permits a comparison of freight rates because the equilibrium rates under monopoly and monopolistic competition are obtained as $g_x((N)_m)$ and $g_x((N)_{mc})$, respectively. This comparison is straightforward for a given situation in which all parameters are known and the function $g_x(N)$ may be constructed. However, it is desired to obtain some generalizations concerning the comparison, which would hold for all numerical values of parameters. It is known that $(N)_m \leq (N)_{mc}$. Therefore knowledge of the shape of $g_x(N)$ would suffice to obtain some general information concerning comparisons between $g_x((N)_m)$ and $g_x((N)_{mc})$. Unfortunately, unlike the case of $\pi_{2x}(N)$, the form of $g_x(N)$ cannot be designated a priori.

There is one situation in which a precise relationship between $g_x((N)_m)$ and $g_x((N)_{mc})$ may be specified. Suppose that $(g_j^B)_m$ maximizes $g_j^B x_j^A$ for all $j = 1, \ldots, u$ and $(g_k^A)_m$ maximizes $g_k^A x_k^B$ for all $k = 1, \ldots, v$. The variable $g_j^B x_j^A$ is total revenue from commodity j net of the commodity costs incurred in its transportation ($g_j^B x_j^A = f_j^B x_j^A - c_j^B x_j^A$) and will be called 'gross profits' (derived from carrying commodity j). A similar remark applies to $g_k^A x_k^B$. Although the situation has been described in terms of aggregate demand functions, it may also be expressed in terms of the demand functions faced by an individual vessel. In fact, if gross profits from any commodity j (or k) are maximized by (or for) each vessel, they are also maximized in toto, and conversely.

Suppose that for every commodity j (or k), $(g_j^B)_m$ (or $(g_k^A)_m$) maximizes gross profits for each vessel. To view this special case in diagrammatic terms, consider Figure 3.4. Although it is concerned with the transportation of a given

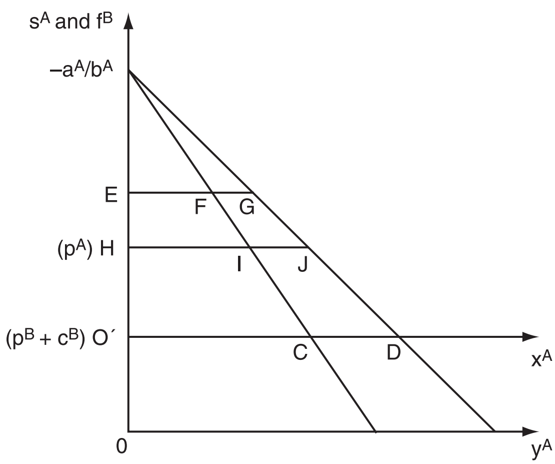

Figure 3.4 Maximization of gross profits.

(non-subscripted) commodity from B to A (as in Section I), the principles derived from it are general. Since the freight rate (f^B) will not be below c_B, the horizontal axis for the transportation demand function extends from $(p^B + c^B)(=OO')$. The independent variable is then g^B rather than f^B. Suppose that the transportation demand function facing each vessel in monopolistic (conference) equilibrium is based on the straight line from $-a^A/b^A$ to D. Consider the point of unit elasticity on this segment. Denote this point by G and from it draw a horizontal line cutting the vertical axis at E. Let H be the point on the vertical axis level with the domestic supply function. Thus $p^A = OH$.

The supposition regarding gross profits implies the following selection of g^B.

(i) $OH \geq OE \Rightarrow g^B = O'E$

(ii) $OO' < OH \leq OE \Rightarrow g^B = O'H$

Case (ii) is portrayed in Figure 3.4. In this situation provision of transportation services is constrained to be in the inelastic portion of the segment (except in the limiting situation in which unit elasticity is achieved). Inelasticity implies maximization of gross profits by restricting output.

It must be emphasized that the phenomenon represented by the supposition is a special circumstance. It is not the general case. Equilibrium under monopoly quite possibly may involve relationships that violate (i) and (ii). This will occur if maximizing gross profits for every commodity creates capacity requirements that exceed the limits of each vessel under $(N)_m$. Increasing the size of the vessel and/or adding sufficient ships to the route in order to reduce capacity requirments per vessel, would enable the achievement of gross-profits maximization for all commodities. However, total net profits (π_1) would be reduced, because the costs of adding capacity (a multiple of $(V + D)$) would exceed the gain in total gross profits. If capacity were costless ($V = D = 0$), then maximization of π_1 would imply, and be implied by, maximization of gross profits for every commodity.

If under monopolistic equilibrium (i) and (ii) do hold for all j and all k, then $(g)_m = (g)_{mc}$. The reasoning is cogent. Assume that the premise is true. There are two cases. Either $(N)_{mc} = (N)_m$ or $(N)_{mc} > (N)_m$. If $(N)_{mc} = (N)_m$, then the same demand functions are faced by each vessel under the two structures. For each vessel, maximizing gross profits for every commodity maximizes total net profits (π_2), because by hypothesis (as the monopolistic equilibrium shows) this can be achieved without violating capacity constraints. If $(N)_{mc} > (N)_m$, the demand functions under monopolistic competition shift inward toward the origin, compared to the corresponding ones under monopoly. Because of the peculiar nature of this shift, namely, rotation about $-a^A/b^A$ (or $-a^B/b^B$), the same freight rates that maximize gross profits under $(N)_m$ maximize them under $(N)_{mc}$. The general proposition is that given a demand curve of constant negative slope, its rotation about the intercept with the vertical axis preserves the value of elasticity among points identified by lying on the same horizontal line. For example, in Figure 3.4, F and G have the same, namely, unitary elasticity. Similarly, I and J have the same elasticity. Because demand functions shift inward under monopolistic competition, the usage of capacity (on a per vessel basis) is less than in monopolistic equilibrium, and in the latter case maximization of gross profits satisfies the capacity constraints.

Finally, it is now apparent that, under monopoly, maximization of aggregate gross profits for a particular commodity occurs if and only if such maximization takes place for each vessel individually. The properties of rotation of a demand function about its intercept with the vertical axis, together with the assumption that an identical demand function faces every vessel, characterize the assertion as obvious.

The case of maximization of total gross profits in both monopoly and monopolistic competition, hence of $(g)_{mc} = (g)_m$, coupled with $(N)_{mc} > (N)_m$, illustrates the 'wastes' of monopolistic competition. The bundles of commodities transported are identical under both market structures: $(x_j^A)_{mc} = (x_j^A)_m$ for all j and $(x_k^B)_{mc} = (x_k^B)_m$ for all k. This always occurs when $(g)_{mc} = (g)_m$. Though cargoes are identical, costs are not. The existence of more vessels under monopolistic competition implies that costs are greater by a multiple of $(V + D)$ and that unused capacity per vessel is greater. It is noteworthy that the extra costs and excess capacity of monopolistic competition are here expressed relative to monopoly rather than pure competition. The wastes are due to the introduction of a competitive element (free entry) into a monopolistic situation (price-making). Interestingly enough, contrary to the implication of my analysis that free entry into the conference would be inefficient relative to a complete conference monopoly, Ferguson et al. seem to take the opposite view. They write: 'U.S. policy has actually forced admission of some carriers that otherwise would have been less effective competitors of the conferences' [2, p. 396].

VI. Summary

In this chapter I developed an approach to the international-transportation industry (with special reference to shipping conferences) which showed that freight-rate determination in this industry is a special case of the peak-load pricing problem. A programming model of profit-maximization under constraints was found to apply to both pure monopoly and monopolistic competition, with the implication that the latter market structure is relatively inefficient. The policy prescription is that to foster free entry of outside liners into conferences might lead to greater inefficiency in the oceanic shipping industry. If this prescription is correct, Ferguson et al. report, in effect, that U.S. shipping policy has pursued the wrong course: 'The consensus is that the U.S. Maritime authorities have compelled conferences to follow a more 'open' admission policy. This implies, of course, that the conferences have power to exclude, if left alone, and that U.S. policy has weakened or eliminated that power' [2, p. 396].

UNCTAD has questioned: 'If the monopolistically determined freight rates of liner conferences were to give way to a more competitive rate structure, how might this affect the trade of developing countries?' It concluded: 'Available knowledge of conference working is perhaps insufficient to allow dogmatic views about the effects of an undermining of conference power' [15, p. 18]. My analysis again suggests that the better policy would be to foster competition outside the conferences themselves rather than to press them for a more liberal policy on the entry of new members.

Appendix

To prove that the transportation demand function is specified by relationships (1) to (9), inclusive, one must pursue all the implications for the demand function provided by each relationship in turn. To be useful, each relationship must provide at least one such implication that represents a genuine step in the development of the function. (Such implications are marked with an asterisk.) Furthermore, all other implications of the relationship must be consistent with the function.

$$* \; x_A \geq 0 \qquad (1)$$

The nonsense result $x^A < 0$ is excluded.

$$x^A(x^A - R^A) = 0 \qquad (2)$$

$$* \; R^A < 0 \Rightarrow x^A = 0$$

$$* \; R^A = 0 \Rightarrow x^A = 0$$

$$R^A > 0 \Rightarrow x^A \geq 0$$

$$Q^B x^A \leq 0 \qquad (3)$$

$$Q^B < 0 \Rightarrow x^A \geq 0$$

$$Q^B = 0 \Rightarrow x^A \gtreqless 0$$

$$* \; Q^B > 0 \Rightarrow x^A \leq 0 \Rightarrow x^A = 0$$

$$(z^B - Q^B)(z^B - L^B) = 0 \qquad (4)$$

$$z^B - Q^B \geq 0$$

$$z^B - L^B \leq 0$$

The above 'implications' of (4) constitute the properties of z^B.

$$R^A(z^B - Q^B) \leq 0 \qquad (5)$$

$$R^A < 0 \Rightarrow z^B - Q^B \geq 0$$

$R^A = 0 \Rightarrow z^B - Q^B \gtreqless 0$

* $R^A > 0 \Rightarrow z^B - Q^B \leq 0 \Rightarrow z^B - Q^B = 0$

$$R^A(z^B - L^B) \leq 0 \tag{6}$$

* $R^A < 0 \Rightarrow z^{B)} - L^B \geq 0 \Rightarrow z^B - L^B = 0$

$R^A = 0 \Rightarrow z^B - L^B \gtreqless 0$

$R^A > 0 \Rightarrow z_B - L^B \leq 0$

$$R^A\,(x^A - R^A) \leq 0 \tag{7}$$

$$R^A\,z^B\,(x^A - R^A) \leq 0 \tag{8}$$

* $R^A > 0$ and $Q^B < 0 \Rightarrow z^B = Q^B < 0 \Rightarrow R^A z^B < 0 \Rightarrow x^A - R^A = 0$
(≤ 0 by (7) and ≥ 0 by (8)

$$Q^B Q^B L^B + x^A - R^A \geq 0 \tag{9}$$

* $R^A > 0$ and $Q^B = 0 \Rightarrow x^A - R^A \geq 0 \Rightarrow x^A - R^A = 0$

It can be shown that all other implications of relationships (7), (8) and (9) are consistent with the demand function.

Notes

* Originally published in *Western Economic Journal*, vol. IX, no. 2 (June 1971), pp. 134–156 (online as *Economic Inquiry*). Reproduced with permission of Oxford University Press.

1 A notable exception, representing the most systematic treatment of the subject, is Hadley and Kemp [5]. See also Kemp [6] and Woodland [14].

2 See, for example, Lefeber [7].

3 The domination of foreign over domestic supply in the case of equal landed prices in A ($f^B = p^A - p^B$) is specified for convenience in the programming formulation of the model. The reverse domination could be incorporated by then setting p^B slightly above its true value when $p^A - p^B > 0$. Of course, in a general-equilibrium model, such as that of Hadley and Kemp [5], the source of supply of commodities on the brink of importation would not be specified a priori.

4 In the special case $p_i^A = p_i^B$, a commodity i is potentially transportable in neither direction.

5 By definition, a measurement ton is 40 cubic feet.

6 Similar breakdowns of costs are presented by Ferguson et al. [2, pp. 108–64], Goss [3 pp. 111–15] and [4, pp. 75–78], McLachlan [8, pp. 327–28] and Sturmey [12 p. 7].

7 These costs are designated elsewhere as cargo handling expenses [2, p. 108], commission and cargo-handling [4, p. 76], cargo charges [8, p. 327] and voyage variables (cargo handling charges) [12, p. 7].

8 Vessel expenses are described elsewhere as vessel operating expenses, port dues and charge [2, p. 108], fuel, crew wages, port charges [8, p. 327], and voyage overheads [12, p. 7]. Their components are listed in detail in [4, p. 76].

9 These expenses are also designated as vessel depreciation charges and general administration expenses [2, p. 108], repairs, insurance, sundries, and

administration [4, p. 76], depreciation and repairs [8, p. 327] and organization overheads [12, p. 7].

10 Of course, subsidy arrangements are ruled out. The case of compensatory rates, in which freight rates below marginal costs for some commodities are compensated for by rates above marginal costs for other commodities, could be handled within the model by considering the two sets of commodities together as one joint commodity. After all, if the total marginal costs of all the commodities involved in the compensatory rates arrangement are not covered by the carrier's receipts, again pricing policy is irrational.

11 For lucid descriptions of shipping conferences and their history, one may consult Ferguson et al. [2, pp. 343–436] and Sturmey [10, pp. 322–58].

12 Various techniques have been developed by shipping conferences to control entry. These include contract preferences (dual rates), deferred rebates, and 'fighting ships' (the sole function of which is to undercut the freight rates of competitors). Entry may also be restricted by natural barriers, such as high capital requirements.

13 The solution value of N is, in general, $N_c + F$, where N_c is a non-negative integer and F a fraction. Retaining the assumption of fixed vessel size would involve then simply eliminating (14) and solving two further problems, one with the variable N replaced by the predetermined number N_c, the other with it replaced by $N_c + 1$. The solution with the higher value of π_1 would be the solution to the original problem.

14 If the vessels of the respective conference members differ in efficiency, the problem of allocation of cargo is not so easily solved. On the one hand, profit-maximization dictates a greater share of cargo allocated to the lower-cost firms and vessels. On the other hand, the bargaining process within the conference might operate to provide more cargo to the less-efficient firms than a single-firm monopoly would allocate, and profit-maximization of the cartel would be curtailed; or, the less-efficient firms might yield cargo in return for a favorable revenue-pooling arrangement.

15 Ferguson et al. confuse the issue by suggesting that nondiscriminatory rates would be based on costs rather than capacity utilization. 'The *essence* of discrimination in the economic sense is charging different buyers what they can be made to "bear," rather than prices closely associated with costs' [2, p. 356]. Again, '. . . rates, given the conference system, . . . are discriminatory in the sense that differences in them do not reflect differences in marginal cost' [2, p. 433]. The same mistake is made by UNCTAD: 'Discrimination in this sense refers to differences in prices which are not explained by differences in cost' [15, p. 8]. Such assertions amount to discrimination by definition. Vessel and overhead expenses cannot be allocated to commodities on the basis of costs, as these expenses are not marginal to the carriage of cargo.

16 The peak-load pricing problem is the subject of an extensive literature. A comprehensive bibliography is provided by Pressman [9, pp. 325–26].

17 While it is possible for a vessel to carry no cargo at all in one direction or to carry it at freight rates equal to only commodity costs, it will not continue to service the route on a long-run basis unless the cargo transported in the other direction (and, in general, those carried in both directions together) at least covers vessel and overhead expenses. In this article I do not discuss such matters as waiting in port for extra cargo or laying-up vessels, but these issues are analyzed elsewhere, such as by Goss [4] and Sturmey [11, pp. 191–95].

18 'The monopolistic position occupied by most liner shipping companies makes it theoretically possible for them to demand even higher freight rates, but as no protection exists against competition by newly-established lines, and such competition usually leads to devastating economic results, liner shipping companies are usually compelled to exercise restraint' (Thorburn [13, pp. 125–26]). 'At least at

some rate levels, non-conference services, including those of tramp ships, may become effective substitutes for conference liner service' (Ferguson et al. [2, p. 356]).

19 'The object of conference arrangements is to eliminate competition on price because such competition would force all rates to the level of direct costs whenever a surplus of tonnage appeared' (Sturmey [10, p. 327]). 'When cargoes are scarce, however, long run profitability becomes synonymous with immediate survival, and the attitude of mutual accommodation within the conference is replaced by a more appropriate spirit of individualism: in a severe rate war, not even prime costs (loading and discharging expenses) provide a lower limit to the fall in rates' (McLachlan [8, pp. 334–35]). 'There is therefore hardly any limit to the decrease in freight rates when competition is resorted to between liner shipping firms. . . . As the number of ship-owners calling at the same or adjacent ports is often very small, and all are conscious of the grave economic results of unbridled competition, it is only natural that agreements in so-called conferences are the rule all over the world' (Thorburn [13, pp. 118–19]).

References

1. Abrahamsson, B. J., 'A Model of Liner Price Setting,' *Jour. Transport Econ. Pol.*, Sept. 1968, **2**, 321–31.
2. Ferguson, A. R., et al., *The Economic Value of the United States Merchant Marine*. Evanston 1961.
3. Goss, R. O., 'Investment in Shipping and the Balance of Payments,' *Jour. Indus. Econ.*, Mar. 1965, **13**, 103–15.
4. —— 'The Turnaround of Cargo Liners and its Effect upon Sea Transport Costs,' *Jour. Transport Econ. Pol.*, Jan. 1967, **1**, 75–89.
5. Hadley, G. and M. C. Kemp, 'Equilibrium and Efficiency in International Trade,' *Metroeconomica*, May–Aug. 1966, **18**, 125–41.
6. Kemp, M. C., *The Pure Theory of International Trade*. Englewood Cliffs 1964.
7. Lefeber, L., *Allocation in Space: Production, Transport and Industrial Location*. Amsterdam 1958.
8. McLachlan, D. L., 'The Price Policy of Liner Conferences,' *Scottish Journal of Political Economy*, Nov. 1963, **10**, 322–35.
9. Pressman, I., 'A Mathematical Formulation of the Peak-Load Pricing Problem,' *Bell Journal of Economics and Management Science*, Autumn 1970, **1**, 304–26.
10. Sturmey, S. G., *British Shipping and World Competition*. London 1962.
11. —— 'Economics and International Liner Services,' *Jour. Transport Econ. Pol.*, May 1967, **1**, 190–203.
12. —— *Some Aspects of Ocean Liner Economics*. Manchester 1964.
13. Thorburn, T., *Supply and Demand of Water Transport*. Stockholm 1960.
14. Woodland, A. D., 'Transportation in International Trade,' *Metroeconomica*, May–Aug. 1968, **20**, 130–35.
15. UNCTAD, *Shipping and the World Economy*. New York 1966.

4 Discrimination in the international transportation industry*

In a recent article [3] I examined the conditions for international-transportation equilibrium under the alternative market structures of monopoly (the conference system) and monopolistic competition (a breakdown of the conference). The model involved the assumptions of constant costs of production, linear downward-sloping commodity demand functions, pure competition in commodity markets, two countries represented by discrete spatial points (A and B), constant per-unit commodity costs of transportation (costs allocable to particular cargo for carriage in a given direction – either A to B or B to A), a unique ship type with given vessel and overhead expenses (joint costs of a round-trip A–B voyage), and identical cargo contents of all vessels on the route. In the present article I explore the implications of the model for the analysis of freight-rate discrimination in oceanic transportation.

Consider the following notation (all from [3]).

u = number of commodities potentially transportable from B to A, a typical such commodity denoted as j

v = number of commodities potentially transportable from A to B, a typical such commodity denoted as k

a_j^A (a_k^B) = quantity-axis intercept of the demand function for commodity j (k) in country A (B)

b_j^A (b_k^B) = slope of the demand function for commodity j (k) in country A (B)

p_j^A (p_k^B) = unit cost of production of commodity j (k) in country A (B)

x_j^A (x_k^B) = number of units of commodity j (k) transported from B to A (A to B)

c_j^B (c_k^A) = commodity costs per unit of commodity j (k) transported from B to A (A to B)

f_j^B (f_k^A) = unit freight rate for transportation of commodity j (k) from B to A (A to B)

g_j^B (g_k^A) = unit freight rate net of commodity costs for transportation of commodity j (k) from B to A (A to B)

w_j (w_k) = number of weight tons per unit of commodity j (k)

m_j (m_k) = number of measurement tons[1] per unit of commodity j (k)

W = maximum number of weight tons of cargo that can be carried by the given ship type

M = maximum number of measurement tons of cargo that can be carried by the given ship type

V = vessel expenses of an A–B voyage

D = overhead expenses of an A–B voyage

N = number of ships on the A–B route

π_1 = profits of all ships on an A–B voyage

π_2 = profits of any one ship on an A–B voyage

Then equilibrium under monopoly and monopolistic competition, analyzed in detail elsewhere [3], may be summarized. Let the function $\pi_{2x}(N)$ be constructed as follows. For any given positive N select the freight rates g_j^B, $j = 1, \ldots, u$ and g_k^A, $k = 1, \ldots, v$ to maximize per-vessel profits

$$\pi_2 = \sum_j g_j^B x_j^A/N + \sum_k g_k^A x_k^B/N - (V + D) \tag{1}$$

subject to demand conditions,[2]

$$x_j^A = \begin{cases} a_j^A + b_j^A (p_j^B + f_j^B) \text{ for } f_j^B + p_j^B - p_j^A \le 0 \text{ and} \\ \qquad\qquad a_j^A + b_j^A(p_j^B + f_j^B) \ge 0 \qquad j = 1, \ldots, u \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

$$x_k^B = \begin{cases} a_k^B + b_k^B (p_k^A + f_k^A) \text{ for } f_k^A + p_k^A - p_k^B \le 0 \text{ and} \\ \qquad\qquad a_k^B + b_k^B (p_k^A + f_k^A) \ge 0 \qquad k = 1, \ldots, v \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

capacity constraints,

$$\sum_j w_j x_j^A \leq NW, \sum_j m_j x_j^A \leq NM \tag{4}$$

$$\sum_k w_k x_k^B \leq NW, \sum_k m_k x_k^B \leq NM \tag{5}$$

and freight-rate identities

$$f_j^B = c_j^B + g_j^B \qquad j = 1, \ldots, u \tag{6}$$

$$f_k^A = c_k^A + g_k^A \qquad k = 1, \ldots, v \tag{7}$$

The conference (monopoly) selects that N on $\pi_{2x}(N)$ which maximizes fleet profits

$$\pi_1 = N\pi_{2x} \tag{8}$$

while under monopolistic competition free entry involves the minimization of per-vessel profits, π_{2x}, subject to $\pi_{2x} \geq 0$ and $N \geq 0$.[3]

I. Definition of non-discriminatory freight rates

Both monopoly and monopolistic competition entail pricing (selection of the freight rates g_j^B, $j = 1, \ldots, u$ and g_k^A, $k = 1, \ldots, v$) based on the $u + v$ *individual* transportation demand functions (2) and (3). Aggregate demand functions are not constructed in order to obtain market equilibrium. Although the provision of transportation is an intermediate good (rather, service), consider the analogy of demand for a final commodity. The demand functions of individual consumers of a particular commodity are added to obtain a market demand function; and pricing, irrespective of the market structure, generally is based on aggregate demand – not on specific conditions of individual demand functions. Otherwise there is price discrimination. Similarly, non-discriminatory pricing of oceanic transportation services would entail no attention to individual demands for carriage, except insofar as they affect aggregate demand. Non-discrimination should entail provision of services at a common rate irrespective of the commodity characterization of freight.

The transportation demand functions, in the form (2) and (3), cannot be aggregated. This is because demand is expressed in units of the particular commodity. In general, it is impossible to aggregate x_1 units of commodity 1 and x_2 units of commodity 2 into $(x_1 + x_2)$ 'units' of 'commodity.' The dimensions of x_1 and x_2 differ.

The allocable marginal costs (c_j^B and c_k^A) are associated with expenses of loading and unloading, special costs of handling, etc. They are not a charge for occupancy of capacity per se. Then what are the vessels selling, apart from services allocable to commodity costs? Capacity for a particular direction on a fixed voyage. What are the dimensions of this capacity? Weight and volume. What is the total cost (on a per vessel basis)? There are two components: *D* for the vessel itself, for its existence on a long-run basis, including its replacement when depreciated, and *V* for the voyage.

The above analysis of costs is useful for determining what services the vessels actually sell and for re-interpreting (6) and (7) in this new light. Commodity costs are for services pertaining to a particular commodity; these services are provided at cost. The remaining component of a freight rate is the charge for utilization of capacity. In equations (6) and (7) c_j^B (c_k^A) is the charge for allocable services and g_j^B (g_k^A) the charge for capacity. Since c_j^B (c_k^A) is known, the problem is to determine g_j^B (g_k^A). It is important to realize the difference between this interpretation and the one expressed in [3, p. 140]. The latter is of general applicability. It divides the typical freight rate (f_i) into known (c_i) and unknown (g_i) parts on the basis of rationality. It is inconceivable that a freight rate be less than commodity cost; this holds irrespective of the market structure. Thus the sole problem in price determination is to obtain g_i. The new approach considers the commodity cost (c_i) as pertaining to special services for a particular commodity, services that are provided at cost. The g_i component of the freight rate is a charge for *joint* services: provision of capacity. By definition, joint services cannot be allocated to particular commodities on the basis of cost. Therefore it is only logical that any contribution to profits also be included in g_i, justifying the provision of special commodity services at marginal cost (c_i) under any market structure.

Under these circumstances can a non-discriminatory structure of freight rates exist? The existing literature on oceanic transportation asserts, in effect, that *any* pricing system is inherently discriminatory. Consider the following statements: '. . . rates, given the conference system, . . . are discriminatory in the sense that differences in them do not reflect differences in marginal cost' [2, p. 433]. 'Discrimination in this sense refers to differences in prices which are not explained by differences in cost' [7, p. 8]. If the conventional wisdom is that non-discriminatory rates must be based on costs, then all pricing systems are discriminatory. Vessel and overhead expenses are joint costs; as such they cannot be allocated to individual commodities on a marginal-cost basis.

The solution to this dilemma is to define a non-discriminatory freight-rate structure as one based on *capacity utilization* rather than costs. As argued above, g_i may be interpreted as a charge for the service of providing capacity for commodity *i*. Then the absence of discrimination would imply that all commodities be assessed the same charge. However, there are two dimensions of capacity (weight and volume) and two directions of movement (B to A and A to B). Thus there is a total of four freight rates for occupancy of capacity, namely:

r^B (r^A) = freight rate per weight ton for transportation from B to A (A to B)

q^B (q^A) = freight rate per measurement ton for transportation from B to A (A to B)

A non-discriminatory pricing structure must apply a unique set of these four freight rates, irrespective of the commodity transported. Then non-discriminatory commodity rates for the use of capacity may be expressed in terms of these underlying occupancy rates as follows:

$$g_j^B = r^B w_j + q^B m_j \qquad j = 1, \ldots, u \tag{9}$$

$$g_k^A = r^A w_k + q^A m_k \qquad k = 1, \ldots, v \tag{10}$$

To test whether this definition of non-discrimination is valid, one should check that such a freight-rate structure is based on aggregate rather than individual demand functions. Can aggregate demand functions for transportation services be constructed undér a freight-rate structure given by (9) and (10)?

The transportation demand functions (2) and (3) may be re-expressed as follows, for the given values of the parameters a_j^A, b_j^A, p_j^B, $j = 1, \ldots, u$; and a_k^B, b_k^B, p_k^A, $k = 1, \ldots, v$:

$$x_j^A = f_{1j}^A\,(g_j^B) \qquad j = 1, \ldots, u \tag{11}$$

$$x_k^B = f_{1k}^B\,(g_k^A) \qquad k = 1, \ldots, v \tag{12}$$

Substituting (9) and (10) into (11) and (12), respectively,

$$x_j^A = f_{1j}^A\,(r^B w_j + q^B m_j) \qquad j = 1, \ldots, u \tag{13}$$

$$x_k^B = f_{1k}^B\,(r^A w_k + q^A m_k) \qquad k = 1, \ldots, v \tag{14}$$

Because w_j, m_j, $j = 1, \ldots, u$ and w_k, m_k, $k = 1, \ldots, v$ are given parameters, the functions (13) and (14) may be re-expressed with a two-dimensional vector as the price variable:

$$x_j^A = f_{2j}^A\,(r^B, q^B) \qquad j = 1, \ldots, u \tag{15}$$

$$x_k^B = f_{2k}^B\,(r^A, q^A) \qquad k = 1, \ldots, v \tag{16}$$

Consider the following notation:

T_j^A (T_k^B) = utilization of weight capacity by commodity j (k) transported from B to A (A to B), in weight tons

U_j^A (U_k^B) = utilization of volume capacity by commodity j (k) transported from B to A (A to B), in measurement tons

The transportation demand functions (15) and (16) may be altered to refer to the demand for capacity. Because capacity has both a weight and a volume dimension, the quantity variable becomes a two-dimensional vector.

$$(T_j^A, U_j^A) = f_{3j}^A (r^B, q^B) \qquad j = 1, \ldots, u \tag{17}$$

$$(T_k^B, U_k^B) = f_{3k}^B (r^A, q^A) \qquad k = 1, \ldots, v \tag{18}$$

These functions are defined as follows:

$$T_j^A = w_j f_{2j}^A (r^B, q^B),\ U_j^A = m_j f_{2j}^A (r^B, q^B) \qquad j = 1, \ldots, u \tag{19}$$

$$T_k^B = w_k f_{2k}^B (r^A, q^A),\ U_k^B = m_k f_{2k}^B (r^A, q^A) \qquad k = 1, \ldots, v \tag{20}$$

The functions (17) and (18) each possess important properties of comparability, unlike the corresponding functions (2) and (3). The dependent variable has the same dimensions, i.e., the same units of measurement, over all commodities. Furthermore, the independent variable is identical for all commodities. These two properties imply that aggregate demand functions may be constructed, one for each direction – B to A, aggregating (17), and A to B, aggregating (18). Let

$T^A(T^B)$ = total demand for weight capacity on the part of commodities transported from B to A (A to B), in weight tons

$U^A(U^B)$ = total demand for volume capacity on the part of commodities transported from B to A (A to B), in measurement tons

The aggregate demand functions are:

$$(T^A, U^A) = f_4^A (r^B, q^B) \tag{21}$$

$$(T^B, U^B) = f_4^B (r^A, q^A) \tag{22}$$

where (21) is the total demand function for transportation from B to A, obtained by aggregating the individual demand functions (17) as follows. Let $(\bar{r}^B, \bar{q}^B)$ be a given value of (r^B, q^B). For $j = 1, \ldots, u$, $(\overline{T}_j^A, \overline{U}_j^A) = f_{3j}^A(\bar{r}^B, \bar{q}^B)$. Then $\overline{T}^A = \sum_j \overline{T}_j^A$ and $\overline{U}^A = \sum_j \overline{U}_j^A$ are the values of T^A and U^A corresponding to $(\bar{r}^B, \bar{q}^B)$. This procedure can be followed for all values of (r^B, q^B). In other words, vector addition is applied for each value of (r^B, q^B):

$$(T^A, U^A) = (\sum_j T_j^A, \sum_j U_j^A) = \sum_j (T_j^A, U_j^A)$$

$$= \sum_j f_{3j}^A (r^B, q^B) = f_4^A (r^B, q^B)$$

The total demand function for transportation from B to A, (22), is obtained by aggregating the individual demand functions (18) in a similar manner. The summing of each set of functions, (17) and (18), is analogous to the conventional addition of individual demand curves. The special feature here is that both price and quantity demanded are two-dimensional rather than uni-dimensional. Thus I have shown that aggregate demand functions, one for each direction of transportation, underlie a non-discriminatory pricing scheme based on the occupancy of capacity.

II. Monopoly and monopolistic competition under non-discriminatory pricing

Define the following freight rates as holding net of commodity costs:

r_j^B (r_k^A) = freight rate per weight ton for transportation of commodity j (k) from B to A (A to B)

q_j^B (q_k^A) = freight rate per measurement ton for transportation of commodity j (k) from B to A (A to B)

Then, quite generally, the commodity freight rates $g_j^B, j = 1, \ldots, u$ and g_k^A, $k = 1, \ldots, v$ may be expressed as follows:

$$g_j^B = r_j^B w_j + q_j^B m_j \qquad j = 1, \ldots, u \tag{23}$$

$$g_k^A = r_k^A w_k + q_k^A m_k \qquad k = 1, \ldots, v \tag{24}$$

The rates $r_j^B, q_j^B, j = 1, \ldots, u$ and $r_k^A, q_k^A, k = 1, \ldots, v$ need not be considered prices for occupancy of capacity. Equations (23) and (24) can apply to any market structure and any pricing policy. Given an arbitrary value of g_j^B (g_k^A), at least one value of the vector (r_j^B, q_j^B) $((r_k^A, q_k^A))$ can be found that satisfies (23) ((24)).[4] Indeed, the problem of maximizing (1) subject to (2)–(7), the solutions of which yield equilibrium under monopoly and monopolistic competition, may be altered as follows with no effect on results. Replace g_j^B, $j = 1, \ldots, u$ and $g_k^A, k = 1, \ldots, v$ as choice variables by $r_j^B, q_j^B, j = 1, \ldots, u$ and $r_k^A, q_k^A, k = 1, \ldots, v$ and in (1), (6), and (7) by their expressions in terms of the new choice variables, i.e., the right-hand sides of (23) and (24), respectively.

Now suppose that the following constraints are added to the new programming problem:

$$r_j^B = r^B,\ q_j^B = q^B \qquad j = 1, \ldots, u \tag{25}$$

$$r_k^A = r^A,\ q_k^A = q^A \qquad k = 1, \ldots, v \tag{26}$$

Then the resulting pricing scheme must be non-discriminatory and the programming problem reduces to the following. For a given N select the freight rates r^B, q^B, r^A and q^A to maximize per-vessel profits

$$\pi_2 = (r^B T^A + q^B U^A + r^A T^B + q^A U^B)/N - (V + D) \tag{27}$$

subject to the demand conditions, (21) and (22), and the capacity constraints

$$T^A \leq NW,\ U^A \leq NM \tag{28}$$

$$T^B \leq NW,\ U^B \leq NM \tag{29}$$

Of course, the freight rates per unit of commodity would be obtained by application of the identities (6), (7), (9), and (10).

Again, the function π_{2x} (N) is obtained by solving the problem for all positive values of N, and equilibrium under monopoly and monopolistic competition may be determined. In the case of the conference-monopoly, the point selected on π_{2x} (N) is that N which maximizes (8), the fleet profits. With the conference system replaced by monopolistic competition, the equilibrium N is that which minimizes π_{2x} subject to $\pi_{2x} \geq 0$ and $N \geq 0$. The equilibrium freight rates under either circumstances are non-discriminatory. Thus these market structures may be designated as 'non-discriminatory monopoly' and 'non-discriminatory monopolistic competition,' respectively.

Could such market forms ever arise in practice? Essentially, they involve the imposition of constraints (25) and (26) on the per-vessel profit-maximization process. These constraints could be imposed as a matter of policy by a government or by intergovernmental agreement. Alternatively, they could reflect irrational or customary pricing on the part of carriers. In any event, non-discriminatory monopoly and monopolistic competition are of interest as market structures that yield non-discriminatory freight rates, and their equilibria can be compared with those yielded by the corresponding 'discriminatory' market structures, namely, monopoly and monopolistic competition.

For any given N, the set of feasible prices (that is, prices that satisfy the constraints) for the non-discriminatory market structures must be contained within the set of feasible prices for the discriminatory structures. Therefore maximum per-vessel profits under non-discriminatory monopoly and monopolistic competition must be less than (or, at the limit, equal to) the profits under monopoly and monopolistic competition, again for any given number of ships on the route. The π_{2x} (N) function for the non-discriminatory market structures is uniformly below that for the discriminatory structures, except if there exists a value of N for which (25) and (26) are non-binding, i.e., satisfied even though not present as constraints. Such a situation is conceivable but highly unlikely.

It has been shown elsewhere [3, pp. 145–47] that π_{2x} (N) is a monotonically-decreasing function of N. Then one would expect the following relationships among equilibrium per-vessel profits (π_{2x}) and equilibrium number of ships

(N) under the four market forms: monopoly (m), monopolistic competition (c), non-discriminatory monopoly (nm), and non-discriminatory monopolistic competition (nc). In the case of a 'greater (less) than or equal' relationship, the equality sign holds only in the unlikely event that constraints (25) and (26) are non-binding.

(i) $(\pi_{2x})_m \geq (\pi_{2x})_{nm} > (\pi_{2x})_c$, $(\pi_{2x})_{nc}$, where $(\pi_{2x})_c$ and $(\pi_{2x})_{nc}$ can be non-zero (positive) only because of the discrete nature of N.

(ii) $(N)_{nc} \leq (N)_c > (N)_m \geq (N)_{nm} < (N)_{nc}$

III. The problem of freight-rate disparity

It has been alleged that the United States is the victim of discriminatory freight rates in its international trade. This controversy entered the public domain in the mid-1960's in a set of hearings on discriminatory ocean freight rates before the Joint Economic Committee [6]. The concern of the Committee was the hypothesis that U.S. exports are the victims of significantly higher ocean transportation rates than those that apply to exports of other countries. The Committee considered two sub-hypotheses, namely, (1) that U.S. exports face higher freight rates than U.S. imports, and (2) that U.S. exports suffer higher freight rates than foreign exports when the common destination is a third country. However, (1) was the principal subject of its investigations.

The testimony heard by the Committee, the studies presented to it, and the recommendations of that body were marked by the lack of a sound, fundamental, and generally accepted theoretical basis. This was no fault of the Committee, its staff, or the witnesses that appeared before it. It was rather a reflection of a deficiency of the state of international transportation theory; and hence economic theoreticians in this area must bear the blame for having failed to provide guidance. However, Bennathan and Walters, in a recent study of ocean freight rates [1, Ch. 5], apply a theoretical model to a re-examination of the discrimination controversy. They confine themselves to one particular route (the North Atlantic route from the U.S. to Britain and Ireland). They acknowledge that freight rates are higher on the outbound direction of this route (U.S. to Britain and Ireland) than on the inbound direction (Britain and Ireland to U.S.). They proceed to explain the disparity in terms of differential liner-tramp competition on the two legs of the route. In the inbound direction liners (the conference vessels) are under relatively intense competition from tramps (non-conference vessels) compared to that in the outbound direction. So the conference cannot set prices as near to profit-maximizing levels on the inbound direction of the route (i.e., for U.S. imports) as it does on the outbound direction (i.e., for U.S. exports).

Bennathan and Walters might very well be right in their explanation of the freight-rate disparity on the two directions of the particular route they considered. Even so, they miss the main point – and perhaps because of their assumption that the outbound and inbound conferences do not coalesce (or 'collude,' as they say) to form one 'superconference' on the route.[5] In contrast to the Bennathan-Walters study, the model that I have proposed assumes that if there is monopoly pricing of transportation services, it takes the form effectively of a single conference acting as a cartel for both directions on the route. While I agree with Bennathan and Walters that whether the inbound and outbound conferences collude is an empirical question, the identical (or near identical) membership of liner companies in each conference strongly suggests that the two individual conferences, one for each direction of the route, in essence behave as one conference. This conclusion is especially reinforced if one assumes profit-maximizing behavior on the part of a conference (the 'conference-cartel' or 'conference-monopolist'), as both Bennathan-Walters and I do. It is hard to believe that liner companies somehow set freight rates to maximize their total fleet profits on one direction of the route independent of this maximization process in the other direction. Indeed, the very point that Bennathan and Walters neglect to mention – the existence of complementary demands and joint costs – implies that such independent maximization behavior is impossible.

The main feature of my model is that it considers freight-rate determination as a case of peak-load pricing.[6] 'The cargo on the B-to-A part of the round-trip voyage in my model does not compete for space with the cargo on the A-to-B part; yet each has the potential of contributing to the joint costs of the voyage, namely, vessel and overhead expenses. This is a case of complementary demands or what has been termed the peak-load pricing problem' [3, p. 144]. Now, discrimination is a concept that can apply only to demand functions that are competitive. Such functions can be added together to produce an aggregate demand function, and a price structure is non-discriminatory if and only if it is based on the aggregate function. Furthermore, the individual functions are added 'horizontally': for a given price, add up the quantities demanded across all individual functions. Thus the aggregate demand function (21) is obtained by summing the u individual functions (17) in just this manner: for a given price vector, (r^B, q^B), add up the demands for capacity, (T_j^A, U_j^A), across all individual functions $j = 1, \ldots, u$. The addition is performed in this fashion because the u functions (17) are *competitive*, not complementary; they compete for space on the B-to-A route of the voyage. And the aggregate demand function (22) is obtained by adding the v functions (18) in a similar manner.

Thus freight rates are discriminatory unless they take the form (9) and (10), i.e., unless constraints (25) and (26) apply. To speak of 'discrimination' between rates on the A-to-B and those on the B-to-A directions of the route is to assert that a non-discriminatory freight-rate structure must satisfy the following additional constraints:

$$r^B = r^A \tag{30}$$

$$q^B = q^A \tag{31}$$

If constraints (30) and (31) do not hold, then one might assert that a disparity exists between freight rates on the two legs of the voyage. However, such a disparity can in no way be construed as a case of price discrimination. The aggregate demand functions (21) and (22) are complementary, not competitive; demand for capacity in the B-to-A direction is not competitive with demand for capacity in the A-to-B direction. Complementary demand functions are added 'vertically' rather than 'horizontally': for a given quantity supplied, add up the demand prices of the complementary demands satisfied by this supply.

In the case of the complementary functions (21) and (22), their two-dimensional nature does not vitiate the addition of 'prices' for a given 'quantity supplied.' The capacity supplied by N vessels is (NW, NM). If demand is not satiated in either direction by this capacity, then $(T^A, U^A) = (T^B, U^B) = (NW, NM)$ and total fleet revenue (net of commodity costs) is $(r^B + r^A) NW + (q^B + q^A) NM$, as given by vectoral 'vertical' addition of the demand functions. The principle of vertical addition is not destroyed by the fact that suppliers under non-discriminatory monopoly and monopolistic competition may not select the rates (r^B, q^B) and (r^A, q^A) so as to 'offer' (use up) all their capacity in either or both directions. The principle is preserved by the total-revenue term in the objective function (27); total fleet revenue (net of commodity costs) is $(r^B T^A + r^A T^B + q^B U^A + q^A U^B)$, while constraints (28) and (29) apply.

In general, the process of 'vertical' addition involves the assignment of unequal prices to the individual demand functions that are complementary; but this price disparity is not discrimination. Indeed, such disparity would be exhibited by the price systems even of a public monopoly that maximizes social welfare, as first analyzed by Steiner [5], and of a set of suppliers that behave in a purely competitive fashion, as I have pointed out [4].

Therefore the freight-rate disparity that exists between the outbound and inbound legs of any trading route is neither discriminatory nor a reflection of monopoly (conference-cartel) pricing. The Joint Economic Committee would have been advised better to have ignored the will-o'-the-wisp issue of such freight-rate disparity between the directions of a route and to have concentrated on the freight-rate disparity among commodities transported *in the same direction*. The latter disparity is a genuine case of discriminatory pricing and can exist only under a market structure that is not purely competitive.

Notes

* Originally published in *Western Economic Journal*, vol. X, no. 2 (June 1972), pp. 170–181 (online as *Economic Inquiry*). Reproduced with permission of Oxford University Press.

1 A measurement ton is a measure of volume equivalent to 40 cubic feet.

2 These demand conditions are expressed as a set of constraints in [3, p.137].

3 Multiplying (1) by N, one sees that $\pi_{2x} = 0$ when $N = 0$.

4 Actually, for g_j^B (g_k^A) > 0, there are an infinite number of such values of (r_j^B, q_j^B) ((r_k^A, q_k^A)).

5 In fairness to Bennathan and Walters, they do assert that even with a single conference for the route their conclusions probably would remain unchanged.

6 It should be pointed out that the existence of tramp competition is quite consistent with my model. The effect of both tramp vessels and non-conference liners on the pricing policy of the conference (and of its constituent companies in the event of a breakdown of the conference) may easily be incorporated in the programming problem in the form of additional constraints, as outlined in [3, pp. 144–45].

References

1. Bennathan, E. and Walters, A. A. *The Economics of Ocean Freight Rates*. New York 1969.
2. Ferguson, A. R., et al., *The Economic Value of the United States Merchant Marine*. Evanston 1961.
3. Officer, L. H., 'Monopoly and Monopolistic Competition in the International Transportation Industry,' *Western Economic Journal*, June 1971, **9**: 134–56.
4. —— 'The Optimality of Pure Competition in the Capacity Problem,' *Quarterly Journal of Economics*, Nov. 1966, **80**: 647–51.
5. Steiner, P. O., 'Peak Loads and Efficient Pricing,' *Quarterly Journal of Economics*, Nov. 1957, **71**: 585–610.
6. Joint Economic Committee, *Discriminatory Ocean Freight Rates and the Balance of Payments*. Hearings and Reports, 88th and 89th Congresses, 1963–66.
7. UNCTAD, *Shipping and the World Economy*. New York 1966.

5 Afterword to Part I

Chapters 1–4 are a foray into microeconomic theory, focusing on firm and market behavior under conditions of joint supply. A multidimensional approach to pricing is developed. In this 'afterword,' the analyses of the earlier chapters are integrated both among themselves and with the existing literature. Then the issues of 'paper' freight rates, freight-rate discrimination, and freight-rate disparity are given a re-examination. A freight rate is a price, of course; so the treatment can be readily generalized, along the lines of Chapter 2.

Relationship to literature

The genesis of Chapter 1 is the issue of 'optimum' (meaning 'efficient') pricing when demand functions are complementary, that is, involve joint use of capacity. Traditionally, this is called the 'peak-load pricing problem.' The situation is not the same as that of a public good or a common resource; for the good (with respect to either demand – given two demands) has the property of excludability. Also, the situation is not the same as a good made artificially scarce by the supplier or suppliers, because there is a capacity cost that increases with output. In other words, the good does have the property of depletability. The commodity under consideration is a private good produced for both (or all) demands under the condition of joint supply.

The statement that 'capacity cost increases with output' pertains to output for all demands jointly. There is a joint cost – that is 'capacity cost.' The existence of demand-specific costs (called 'operating costs' in the literature) is best handled, following Steiner (1957: 588), by subtracting such costs from the demand price. The implication is that the price for each demand is defined net of specific per-unit operating costs. In particular, a zero optimum price, as in the firm-peak case in Chapter 1, translates into a positive price equal to the demand-specific per-unit operating cost. Steiner, however, does define price to be inclusive of operating cost. He does this by placing the horizontal axis at the level of per-unit operating cost rather than at 'zero' price, as done in Chapter 1. Substantively, the procedure of the authors is the same.

The important result of Chapter 1 is that joint use of capacity by different demands does not generate a market failure. The efficiency of perfect competition under ideal conditions (absence of externalities, perfect information, etc.) asserts itself. Under perfect competition, the sum of consumer surplus and producer surplus continues to be maximized with joint use of capacity. Under constant costs (constant per-unit cost of capacity – Steiner's assumption, retained in Chapter 1), perfect competition continues to maximize consumer surplus subject to producers suffering no monetary loss, that is, subject to zero producer surplus. A monopoly supplier with a linearly homogenous production function corresponds to the production conditions of a perfectly competitive industry of identical firms under constant costs; the actual technology of the firms is irrelevant. As Buchanan (1967: 703) writes: 'Steiner's solution, presented initially as a normative rule for the welfare-maximizing public-enterprise monopoly, is held to describe the equilibrium results of a decentralized competitive process.'

Buchanan observes that the constant per-unit capacity cost can be the result of an increasing-returns-to-scale production function combined with rising supply prices of industry-specific factors. Of course, that is a ' "knife-edged" underpinning of constant costs,' and therefore very unlikely – as commented in a reply to Buchanan (Officer 1967: 705). Nevertheless, Buchanan's main point is well-taken: one must not forget that, with increasing returns to scale, perfect competition would cease to be efficient. That well-known theorem in economics also continues to hold with joint use of capacity! However, the theorem that perfect competition maximizes the sum of consumer surplus and producer surplus remains valid under decreasing returns to scale for the monopoly firm, which corresponds to varying efficiency of the perfectly competitive firms (per-unit cost of capacity differing among the firms).

The Steiner–Officer model assumes independent demands, linear cost functions (constant per-unit operating costs and capacity cost), and no profit constraint. Pressman (1970) generalizes the model in various ways; he allows for interdependent demands, general cost functions, and a profit constraint. However, although Pressman lists Officer (1966 – here Chapter 1) in his references, he follows Steiner in examining monopoly alone as the market structure.

In contrast, Crew (1969) does follow up on the two market structures examined in Chapter 1: a social-welfare-maximizing ('public-spirited') monopoly and profit-maximizing perfectly competitive firms. When plants are perfectly divisible, Chapter 1 demonstrates that the two market structures lead to the same, efficient, result. When plants are indivisible (that is, yield a fixed capacity or supply), then Chapter 1 shows perfect competition does provide feasible efficiency under that market structure; but in general that efficiency involves lower welfare than if the plants were perfectly divisible. Crew's contribution is the observation that Officer's analysis is not carried far enough; for in this situation the monopoly would produce the same output as if plants were continuously divisible and thus maximize social welfare, albeit with a loss of producer surplus.

Chapter 2, in spite of its later original-publication date, was written before Chapters 3–4 and also logically precedes these chapters. Chapter 2 is within the rubric of characteristics-demand and hedonic-function/pricing models. (An excellent survey of these strands of literature and of the relationship between them is Triplett 1987.) However, there are three main differences between Chapter 2 and this literature.

First, almost all the literature – including the important work of Lancaster (1979, 1991) – makes the assumption that the consumption 'input-output' technology generating the characteristics from a given commodity is identical for all consumers. In fact, the only exception that I could find is the model of Ironmonger, described and contrasted with the Lancaster approach, in Wadman (2000: 62–77). Even that exception may be more apparent than real; for operationally Ironmonger's consumption technology 'becomes the average **W** [technology] faced by all consumers, or the *representative consumer* [italics in original] (Wadman 2000: 77). In contrast, Chapter 2 allows the amount of any characteristic (or attribute) that a demander derives per unit of commodity to differ among demanders.

Second, the literature deals only with the situation in which pricing is based directly on the amount of commodity demanded. Goods – not characteristics – are sold as such. Chapter 2 adopts the opposite pricing approach: pricing is based on the amount of each characteristic received by a consumer. The important assumption underlying that pricing model is that the consumption technology, which varies among consumers, is not only measurable but also known – for each consumer – by both the consumer and the producer(s).

Third, at least the economics component of the literature is concerned with the case of multiple commodities. This enables that literature to consider issues such as consumer multiproduct equilibrium, substitutability among goods, varying product quality, and introduction of new commodities. Chapter 2 treats only the case of a single commodity. Thus the chapter does not provide a general theory of consumer or market behavior. The virtue of the chapter is that it is the underpinning of a useful approach to examine the international-transportation – in particular, oceanic cargo shipping – industry. That examination is the topic of Chapters 3–4, with the underpinning clearer in Chapter 4. The relationship of the characteristics approach of Chapter 2 to the oceanic-transportation model of Chapters 3–4 is apparent: the single commodity (Chapter 2) corresponds to two-dimensional capacity (Chapters 3–4); consumers (Chapter 2) become commodities (Chapters 3–4); and characteristics (Chapter 2) are the weight and volume dimensions of transported commodities (Chapters 3–4).

Chapters 3–4 are also related directly to Chapter 1. Determination of freight rates is an application of peak-load pricing. The (two-dimensional) demand for capacity in the A-to-B direction is complementary to that in the B-to-A direction. Vessel and overhead expenses constitute joint costs of capacity – meaning joint with respect to direction of transportation.

Irrelevance of freight-rate expression

In Chapters 3–4, freight rates at the commodity level are expressed per physical unit of commodity, whatever is this physical unit. Thus g_j^B (g_k^A) = freight rate net of commodity (that is, allocable marginal) costs, per unit of commodity j (k), for transportation of commodity j (k) from B to A (A to B). This (g_j^B or g_k^A) is the freight rate of interest. This expression differs from the practice of carriers in two respects. First, freight rates are inclusive of commodity costs, of course. Nevertheless, rational behavior involves non-negative g_j^B or g_k^A prices. (See the discussion in Chapter 3, note 10.) These prices represent the contribution of commodity j or k to vessel expenses, overhead expenses, and profit. So nothing is lost empirically, and much is gained theoretically, by viewing price behavior in terms of the g_j^B and g_k^A.

Second, carriers do not generally state freight rates in terms of physical unit of a commodity; rather, they express rates per weight ton or per measurement ton. Thus the revenue of the carrier (cost to the shipper), in excess of commodity costs, of shipping, say, x_j^A units of commodity i from B to A is $g_j^B \cdot x_j^A$. Now, the unit of commodity j may be redefined to be either a weight ton or measurement ton. Recall that w_j (w_k) = number of weight tons per unit of commodity j (k), and m_j (m_k) = number of measurement tons per unit of commodity j (k). Then the freight rate corresponding to g_j^B simply becomes g_j^B/w_j or g_j^B/m_j, for weight-ton or measurement-ton pricing; and revenue, at $(g_j^B/w_j)\cdot(w_j \cdot x_j^A)$ or $(g_j^B/m_j)\cdot(m_j \cdot x_j^A)$, is unchanged from $g_j^B \cdot x_j^A$. Of course, the same argument applies to the g_k^A associated with shipping x_k^B units of commodity k from A to B.

Therefore the particular formulation of a rate is quite independent of its rationale. Indeed, for some commodities the natural unit may be nothing but a measure of weight or volume content. This does not imply that the stipulated rates are functions only of occupancy of weight or volume capacity of the vessel. It is possible that these properties are unimportant or even absent factors in the determination of the freight rates.

To clarify the point, consider that a freight rate (g_j^B or g_k^A) may be expressed as derivative from two separate freight rates, imposed on the commodity's utilization of weight and volume capacity of the vessel. Let r_j^B (q_j^B) = freight rate per weight (measurement) ton for transportation of commodity j from B to A, and r_k^A (q_k^A) = freight rate per weight (measurement) ton for transportation of commodity k from A to B. Then

$$g_j^B = r_j^B \cdot w_j + q_j^B \cdot m_j \tag{1}$$

$$g_k^A = r_k^A \cdot w_k + q_k^A \cdot m_k \tag{2}$$

Equations (1) and (2) are equations (23) and (24) in Chapter 4. In these equations, only (w_j, m_j) and (w_k, m_k) are predetermined. A large (indeed, infinite) number of pairs (r_j^B, q_j^B) traces out a given value of g_j^B; and a large number of pairs (r_k^A, q_k^A) traces out a given value of g_k^A. Again one sees that the expression of a rate may be independent of the determination of the rate. Nevertheless, the pairs (r_j^B, q_j^B) and (r_k^A, q_k^A) – even though neither pair is unique – may be construed as the component rates of g_j^B and g_k^A.

It should be noted that, as stated in Section I of Chapter 3, the exchange rate between countries A and B is assumed fixed and equal to unity. This simplification enables the analysis to concentrate on the essentials of oceanic transportation.

Freight-rate discrimination

In terms of equations (1) and (2), a non-discriminatory freight-rate structure involves $r_j^B = r^B$ and $q_j^B = q^B$ for all j, and $r_k^A = r^A$ and $q_k^A = q^A$ for all k. So equations (1) and (2) become

$$g_j^B = r^B \cdot w_j + q^B \cdot m_j \tag{3}$$

$$g_k^A = r^A \cdot w_k + q^A \cdot m_k \tag{4}$$

Equations (3) and (4) correspond to the constraints (25) and (26) in Chapter 4. Non-discrimination implies *at most* four different component freight rates: r^B, q^B, r^A, q^A. In words, these rates are the rate per weight (or measurement) ton for transportation of *any* commodity from B to A (or A to B). Discrimination would exist if any of these four rates varies among commodities. The great advantage of defining a nondiscriminatory rate structure via equations (3) and (4) is that aggregate demand functions – one for each direction of trade – may be constructed. The dependent variable is the two-dimensional (weight, measure) capacity demand; the independent variable is the two-dimensional price vector – (r^B, q^B) or (r^A, q^A), depending on whether trade is in the B-to-A or A-to-B direction.

Definitions of discrimination in the literature are different from the definition presented here. Sturmey (1975: 96) states: 'A rate schedule is discriminatory when it contains rates which cannot be justified on the basis of costs incurred allocated between the goods carried on the basis of some definable, but not necessarily, simple, principle.' He elaborates that a freight-rate formula [for g_j^B or g_k^A] can be based on any characteristic of the commodities; he mentions as examples: weight, measurement, value, elasticity of demand, and commodity-specific costs. If there exists a rate that violates the adopted formula, then that rate is discriminatory. Bromley (1987: 35) sees freight-rate discrimination simply as conventional third-degree price discrimination: 'to separate shippers according to [demand] elasticity characteristics.' That is also the approach of Sjostrom (1992: 207): 'Economic price discrimination occurs when the ratio of freight rates to the marginal cost of shipment differs by commodity, with the ratio being higher the more inelastic the demand for shipping the commodity.'

The definitions in the literature have three disadvantages. First, they are not specifically geared to the peak-load aspect of the pricing of oceanic transportation. Second, they are imprecise. The definition of Sturmey is too loose: discrimination is differential treatment with respect to any formula whatsoever. In contrast, the definition of Bromley and Sjostrom relates exclusively to elasticity of demand, but the number of groups to be distinguished is unstated. Third, these definitions do not provide insight in exploring the issue of freight-rate disparity.

Freight-rate disparity

Freight-rate disparity is discussed in Section III of Chapter 4, but only in the context of nondiscrimination. Assume that a freight-rate structure is nondiscriminatory (equations (3) and (4) applying for all j and k). Then there is also the absence of freight-rate disparity if $r^B = r^A$ $(= r)$ and $q^B = q^A$ $(= q)$. So freight-rate nondisparity, combined with freight-rate nondiscrimination, involves equations (3) and (4) becoming

$$g_j^B = r{\cdot}w_j + q{\cdot}m_j \tag{5}$$

$$g_k^A = r{\cdot}w_k + q{\cdot}m_k \tag{6}$$

So nondisparity combined with nondiscrimination involves *at most* two different component freight rates: r and q. In other words, these rates are the rate per weight (or measurement) ton for transportation of *any* commodity in either direction: B to A, or A to B. Disparity, *under the rubric of nondiscrimination*, would exist if either of these two rates varies among commodities.

The problem to be explored here is measurement of freight-rate disparity *without imposing nondiscrimination*. The natural inclination is to compare g_i^B with g_i^A for all i. However, this comparison makes no sense; for, under the model and using the notation of Chapters 3–4, if one of x_i^A or x_i^B is positive, then the other must be zero. In this situation, one of the rates g_i^B or g_i^A is necessarily a 'paper' (irrelevant) rate. This is easily proved. Let c_i^B (c_i^A) denote commodity costs per unit of commodity i transported from B to A (A to B). $x_i^A > 0$ implies $p_i^B + c_i^B \leq p_i^A$, or $p_i^A \geq p_i^B + c_i^B$, which implies $p_i^A + c_i^A \geq p_i^B + c_i^B + c_i^A$, implying $p_i^A + c_i^A > p_i^B$, resulting in $x_i^B = 0$ and thus the level of g_i^A irrelevant. Similarly, it may be proved that $x_i^B > 0$ involves $x_i^A = 0$ and thus the level of g_i^B irrelevant.

An underlying assumption – both here and in Chapters 3–4 – is that commodities are defined (and the transportation of which are priced – via the g_i variables) narrowly, so that intra-industry (meaning intra-commodity) international trade does not exist for any given commodity. That assumption may not accord with empirical reality; but it is key to understanding market behavior, price discrimination, and price disparity in oceanic transportation.

A broader understanding of paper-versus-actual rates builds on the discussion in Section I of Chapter 3. Commodity i is *potentially* transportable from B to A (A to B) only if $p_i^B + c_i^B \leq p_i^A$ ($p_i^A + c_i^A \leq p_i^B$) – and both inequalities cannot be true simultaneously. Whether a commodity is *actually* transported, of course, depends on the value of g_i^B (for B-to-A movement) or g_i^A (for A-to-B movement). Therefore, if there are n commodities in total, u (indexed by j) is the number of commodities potentially transportable from B to A, v (indexed by k) the number of commodities potentially transportable from A to B, and w is the number of commodities potentially not transportable in either direction. So $n = u + v + w$, with u, v and w mutually exclusive.

In sum, it is always true that *at least one* of the rates g_i^A or g_i^B can be set at an arbitrary non-negative value without affecting the flow of trade (transported goods). *Both* rates have this property, if both $p_i^B + c_i^B > p_i^A$ (implying $x_i^A = 0$) and $p_i^A + c_i^A > p_i^B$ (implying $x_i^B = 0$).

The policy implication is that constraining (legally or otherwise) $g_i^A = g_i^B$ for any or all i in no way affects any pricing policy of the carriers. Then can comparisons of g_i^A and g_i^B have any meaning at all? If the parameters p_i^A, p_i^B, c_i^A, c_i^B do not change, are forever fixed, then such comparisons are meaningless. However, if any of these parameters changes its value, then the direction of trade which was a priori nil may be reversed, or absence of trade may be replaced by some trade in one of the directions. If paper rates are set at high levels, such a change may in fact cut off trade rather than reverse its direction, or maintain zero trade. However, price-making carriers presumably recalculate optimum rate schedules upon changes in the basic data (values of parameters). This applies to both the monopoly and monopolistic-competition markets presented in Chapters 3–4. It is only if rate schedules are to be fixed for periods of time encompassing changes in the basic data that commodity-by-commodity comparisons of rates become meaningful. Even then, at any point in time, the above discussion of paper rates is applicable.

In general, therefore, the comparison of actual directional freight rates would not yield meaningful measures of rate disparity. The interest focuses on rate *structure*; the measures are aggregate, rather than commodity-by-commodity, indicators. Three types of measures are considered: (a) averages of freight rates in relation to some other variable, (b) comparisons of actual and normative rates, and (c) indicators of diverted trade. A measure is computed for each direction of trade, and a significant difference in the calculated magnitudes may be taken to designate a directional disparity in the rate structure. In principle, the measures may be computed empirically; but the discussion here (consistent with the approach in Chapters 3–4) is entirely analytical. In all the measures, j is summed over $1, \ldots, u$; and k over $1, \ldots, v$. *This is an important restriction for the measures to be meaningful.*

There are four type-a measures, an average freight rate with respect to some given variable. These measures are:

1. $\Sigma g_j^B x_j^A / \Sigma w_j x_j^A$ versus $\Sigma g_k^A x_k^B / \Sigma w_k x_k^B$

2. $\Sigma g_j^B x_j^A / \Sigma m_j x_j^A$ versus $\Sigma g_k^A x_k^B / \Sigma m_k x_k^B$
3. $\Sigma g_j^B x_j^A / \Sigma p_j^B x_j^A$ versus $\Sigma g_k^A x_k^B / \Sigma p_k^A x_k^B$
4. $\Sigma g_j^B x_j^A / \Sigma d_j x_j^A$ versus $\Sigma g_k^A x_k^B / \Sigma d_k x_k^B$

where $\quad d_j = p_j^A - p_j^B - c_j^B; \quad d_k = p_k^B - p_k^A - c_k^A$

Each of the measures 1–4 is a ratio with total, direction-specific, net (of commodity costs) freight revenue as the numerator. The denominator in 1 and 2 is the total carriage in weight and measurement tons, respectively. These two measures calculate average freight rates (g_j^B and g_k^A) by measuring commodities in comparable units. Thus measure 1 is the average rate per weight ton and measure 2 the average rate per measurement ton. These measures imply that commodities should be assessed freight charges (g_j^B or g_k^A) according to the amount of capacity they require. The implication is that equations (5) and (6) should be imposed as legal or normative constraints on the pricing policy of the carriers. This means nondisparity combined with nondiscrimination. Since measures 1 and 2 each pay attention to only one of the two dimensions of capacity, it is logical that they be used jointly.

There is a major problem with imposing a nondisparity pricing structure under the rubric of nondiscrimination. The differential characteristics of the aggregate directional demand functions – equations (21) and (22) in Chapter 4 – are ignored. As discussed in Section III of Chapter 4, these demand functions are complementary and therefore added 'vertically' rather than 'horizontally.' Then assignment of unequal prices to complementary demand functions is the usual property even of a public monopoly that maximizes social welfare and of unregulated perfectly competitive suppliers (see Chapter 1). Imposition of nondisparity in pricing, given nondiscrimination in pricing, leads to deviations from the price structure that maximizes the sum of consumer and producer surplus.

Measure 3 presents a third way of aggregating the transportation of different commodities: by value (price, or unit cost of production). The denominator of this measure is the total f.o.b. value of cargo. Thus the measure is the average freight charge (g_j^B or g_k^A) per dollar of f.o.b. value of commodity (j or k).The underlying normative concept is that higher-valued cargo should pay absolutely higher – but relatively identical – rates compared to those assessed lower-valued cargo. The policy implication is that carriers should be compelled to set their rates to satisfy the relationship $g_j^B/p_j^B = g_k^A/p_k^A = C$, for $j = 1, \ldots, u$, and $k = 1, \ldots, v$, where C is a constant to be selected by the carriers. The principal disadvantage of measure 3 is its neglect of capacity utilization, both weight and volume.

The denominator of measure 4, unlike those of 1–3, is not an aggregation of cargo. Thus this measure does not have the interpretation of an average rate per unit (however defined) of cargo. The factor d_j or d_k in a term in the denominator is a limit which the freight rate (g_j^B or g_k^A, as always) can exceed only with the result of zero trade in x_j^A or x_k^B. Thus the measure is the weighted average of the ratios of the actual freight rate to its a priori limit beyond which trade and transportation cease. It may be considered an index of the degree of exploitation of a price-taking position on the part of carriers.

However, the measure is afflicted with some weaknesses. First, it implicitly assumes that profits are maximized by maximizing prices (subject to the constraint that there remains nonzero demand). This is not generally true. Setting the freight rate at the designated limit maximizes gross profit for a given commodity only if the domestic supply curve does not cut the domestic demand curve above the latter's point of unit elasticity, with the horizontal axis at $p_j^B + c_j^B$ or $p_k^A + c_k^A$, as the case may be. Second, it may occur that maximizing total *gross* profits does not maximize overall total profit. On all this, see Section V of Chapter 3.

Third, although d_j or d_k is always an upper bound beyond which trade is nil, it may not be the *least* upper bound to the freight rate. If it happens that p_j^A (p_k^B) is above $-a_j^A/b_j^A$ ($-a_k^B/b_k^B$), then, for d_j (d_k) to remain the least upper bound, p_j^A (p_k^B) must be replaced by $-a_j^A/b_j^A$ ($-a_k^B/b_k^B$) in the definition of d_j (d_k).

Type-b measures involve a comparison of the actual rate structure with a normative rate structure. The latter are the rates that emanate from a market with the same demand and cost conditions but with an 'ideal' property. Several alternative normative market forms can be considered. First is perfect competition. That market structure is examined in detail in Chapters 1–2. The treatment in these chapters is general rather than specific to oceanic transportation. However, the analysis and conclusions stand. With perfect competition, pricing is according to characteristics (weight and volume dimensions of commodities) and is nondiscriminatory. As stated in 'Freight-rate discrimination' above, nondiscriminatory pricing involves two aggregate demand functions for (two-dimensional) capacity, one function for each direction of transportation. The perfectly competitive firms – accepting prices (r^B, q^B) and (r^A, q^A) as parameters – throw the entirety of their capacity onto each (directional) market. With free entry, an efficient outcome (x_j^A and x_k^B; $j = 1, \ldots, v$ and $k = 1, \ldots, w$) results. Consumer surplus is maximized subject to zero monetary losses on the part of producers. Subject to an amendment for plant indivisibility (see 'Relationship to literature' above), the perfectly competitive outcome is identical to that of a social-welfare-maximizing monopoly enterprise.

Alternative normative market structures are nondiscriminatory monopoly and nondiscriminatory monopolistic competition, which involves imposing equations (5) and (6) on these market forms. These normative structures are discussed in Section II of Chapter 4.

These three normative market structures (and indeed any other structure deemed normative) yield sets of normative freight rates G_j^B and G_k^A; $j = 1, \ldots, u$; $k = 1, \ldots, v$. These normative rates may be compared with the actual rates, g_j^B and g_k^A, of any given market form. Measure 5 is the comparison:

5. $\Sigma(g_j^B - G_j^B)\cdot x_j^A$ versus $\Sigma(g_k^A - G_k^A)\cdot x_k^B$

For a given direction of trade, the constructed measure is the excess of actual freight revenue over the freight revenue that would prevail given the prices of the normative market structure but the quantities transported of the actual structure. Note that the comparison always is from the standpoint of disparity, not efficiency. So it is acceptable that the individual $(g_j^B - G_j^B)$ and $(g_k^A - G_k^A)$ be either positive or negative, and the measure is total *net* excess freight charges.

In order to construct measures of type-c, it is necessary to consider explicitly domestic production for domestic consumption. Let

x_j^{AA} = number of units of commodity j produced in country A
x_k^{BB} = number of units of commodity k produced in country B

Of course, the entirety of x_j^{AA} (x_k^{BB}) is consumed in the country, A (B), of production.

It is useful to apply the traditional (Viner 1950: 41–55) criterion of trade diversion, namely that supply does not emanate from its real-cheapest source. Certainly, the real cost of supply includes that of transportation as well as production. Then what is the 'real cost' of transportation? One concept, which is the logical minimum to real cost, is allocable marginal (commodity) cost of transportation. So the excess real cost of domestic production of imported commodity j (k) on the part of country A (B) is d_j (d_k), where, it is recalled, $d_j = p_j^A - p_j^B - c_j^B$ $(d_k = p_k^B - p_k^A - c_k^A)$. Measure 6, then, involves comparing the following values of excess real cost:

6. $\Sigma d_{jj} x_{jj}^{AA}$ versus $\Sigma d_{kk} x_{kk}^{BB}$

where jj (kk) runs over those j (k) for which x_j^A $(x_k^B) > 0$. Note that x_j^A $(x_k^B) > 0$ implies d_j $(d_k) > 0$.

Measure 6 gives excess real cost an upward bias. After all, capacity costs of the carriers are part of the real cost of transportation. Integrating type-b and type-c measures, capacity costs may be allocated to individual commodities according to the rate structure of the normative market form adopted. This suggests measure 7:

7. $\Sigma d_{jj} x_{jj}^{AA}$ versus $\Sigma d_{kk} x_{kk}^{BB}$

where jj (kk) now runs over those j (k) for which x_j^A $(x_k^B) > 0$ *and* $d_j - G_j^B$ $(d_k - G_k^A) > 0$. Note that $d_j - G_j^B = p_j^A - p_j^B - c_j^B - G_j^B$ and $d_k - G_k^A = p_k^B - p_k^A - c_k^A - G_k^A$. Measure 7 differs from 6 only in the more restrictive range of summation. A reason why measure 7 uses d_j and d_k again instead of $d_j - G_j^B$ and $d_k - G_k^A$ is that there is no guarantee that the latter magnitudes are non-negative. Another reason is to facilitate comparison with measure 6.

Measure 7 is complementary to measure 5. Measure 7 is total excess real cost of domestic production, while measure 5 is total excess freight charges over real (that is, normative) cost. These measures suffer from deficiencies that are opposite in nature. Measure 7 presents wasteful domestic production as distinct from excess freight charges. It deals with results rather than causes. Its weakness is that it considers only trade that ought to, but does not, take place. No attention is given to trade that does occur. In contrast, measure 5 incorporates trade that is overly increased or overly decreased. Consider the latter situation. If the excess of the actual over normative rate for a given commodity is so high that trade ceases, then the rate difference – though positive (and indeed large!) – has a weight of zero; for the weight is actual trade. One way to eliminate this deficiency is to construct unweighted averages of excess or deficient rates, as follows:

8. $\Sigma(g_j^B - G_j^B)$ versus $\Sigma(g_k^A - G_k^A)$

However, this measure is unsatisfactory, because it gives equal weight to commodities of diverse importance in trade, to commodities transported in small amounts as well as those transported in large quantities. A better way to correct measure 5 is to retain the weighted-average concept, but to let the weight be the sum of imports and domestic production of the commodity. This yields the following indexes:

9. $\Sigma(g_j^B - G_j^B)\cdot(x_j^A + x_j^{AA})$ versus $\Sigma(g_k^A - G_k^A)\cdot(x_k^B + x_k^{BB})$

The demand conditions of the model in Chapters 3–4 imply that at least one of x_j^A, x_j^{AA} (and at least one of x_k^B, x_k^{BB}) is zero. For each commodity, there is only one source of supply: imports or home production. There is no rationing of capacity: for a given g_j^A (g_k^B), the entirety of demand is accommodated by the carriers. It is possible that allowing for rationing would increase overall profit. This would involve changing the model so that demand functions yields the maximum, not the actual, demand. In that situation, both x_j^A and x_j^{AA} (x_k^B and x_k^{BB}) could be positive.

Measure 9 has the advantage of combining the virtues of measures 5 and 7. It gives weight to all trade, actual or potential (that eliminated by excessive pricing). Also, it does that while assigning different commodities weights proportionate to their importance in trade, whether that trade is achieved or thwarted. Again it should be emphasized that aggregate rate disparity, not economic efficiency, is the subject of measurement. So negative values of $(g_j^B - G_j^B)$ and $(g_k^A - G_k^A)$ count as much as positive values.

In principle, normative freight rates and measures 1–9 could be computed as a programming problem, given the data (values of parameters). Chapters 3 and 4 indicate how the programming problem could be solved.

References

Bromley, R. C. (1987) 'Cartels as Alternatives to Contracts: The Case of Ocean Shipping Conferences', unpublished thesis, University of Cambridge.

Buchanan, J. M. (1967) 'The Optimality of Pure Competition in the Capacity Problem: Comment', *Quarterly Journal of Economics*, **81**: 703–5.

Crew, M. A. (1969) 'The Optimality of Pure Competition in the Capacity Problem: Further Comment', *Quarterly Journal of Economics*, **83**: 341–3.

Lancaster, K. (1979) *Variety, Equity, and Efficiency: Product Variety in an Industrial Society*. New York: Columbia University Press.

—— (1991) *Modern Consumer Theory*. Aldershot: Edward Elgar.

Officer, L. H. (1966) 'The Optimality of Pure Competition in the Capacity Problem', *Quarterly Journal of Economics*, **80**: 647–51.

—— (1967) '[The Optimality of Pure Competition in the Capacity Problem] Reply', *Quarterly Journal of Economics*, **81**: 705–6.

Pressman, I. (1970) 'A Mathematical Formulation of the Peak-Load Pricing Problem', *Bell Journal of Economics and Management Science*, **1**: 304–26.

Sjostrom, W. (1992) 'Price Discrimination by Shipping Conferences', *Logistics and Transportation Review*, **28**: 207–16.

Steiner, P. O. (1957) 'Peak Loads and Efficient Pricing', *Quarterly Journal of Economics*, **71**: 585–610.

Sturmey, S. G. (1975) *Shipping Economics: Collected Papers*. London: Macmillan.

Triplett, J. E. (1987) 'Hedonic Functions and Hedonic Indexes', in J. Eatwell, M. Milgate and P. Newman (eds.) *The New Palgrave Dictionary of Economics*, vol. 2, London: Macmillan.

Viner, J. (1950) *The Customs Union Issue*. New York: Carnegie Endowment for International Peace.

Wadman, W. M. (2000) *Variable Quality in Consumer Theory: Toward a Dynamic Microeconomic Theory of the Consumer*. Armonk, NY: M.E. Sharpe.

Part II

Financing of international organizations

6 Are International Monetary Fund quotas unfavorable to less-developed countries?

A normative historical analysis*

Each member country of the International Monetary Fund (IMF or Fund) is assigned a 'quota', and the issue of differential treatment by the Fund to the disfavor of its less-developed members is closely connected to the country distribution of total Fund quotas.[1] Yet there exists no study of developed versus less-developed country (DC versus LDC) disparity in the allocation of IMF quotas.[2] The investigation of 'quota disparity' by level of development is the subject of the present paper. The approach is historical, covering intensely a quarter-century of the Fund's existence, and involves some positive analysis (how the IMF actually determines quotas) but principally normative analysis (how the IMF ought to determine quotas).

Sections I and II are devoted to positive analysis. Section I delineates what the Fund's Articles specify about quota determination; the amazing feature is their total lack of guidance on the determinants of quota level. Section II explores how the IMF establishes quotas. The Fund's stated criterion is *economic size*; but in assigning quotas there is only limited adherence to its own concept of size.

Sections III to VI are concerned with normative analysis. Section III argues that, while the Fund's criterion of economic size may be acceptable in principle, the IMF does a poor job measuring economic size and *prima facie* to the detriment of LDCs. To measure economic size appropriately, components of size must be justified in terms of a mapping from the purposes of quotas to the ultimate criteria for quota determination. Section IV outlines a methodology for evaluating the DC versus LDC distribution of Fund quotas over time using an appropriate measure of economic size. Two levels of analysis are employed: in Section V an aggregate level in which the developed and less-developed members of the Fund are each grouped into one super country, and in Section VI a disaggregate level in which the units of observation are individual countries. A 'quota disparity' is defined as an actual quota that deviates from the quota level that would be predicted on the basis of appropriately measured economic size.

Section VII presents the conclusions of the study, followed by a data appendix. The principal finding is that while quota disparities were in favor of DCs to the detriment of LDCs early in the Fund's history, over time this favoritism declined and according to some tests actually switched to the advantage of LDCs.

I. Quota determination in the IMF articles

The Articles of Agreement of the IMF are explicit that each member is to be assigned a quota, and that this quota is adjustable (but only with the consent of the member concerned). Originally denominated in US dollars, quotas were re-expressed in terms of special drawing rights (SDRs) in March 1972 at a one-to-one ratio of SDRs for dollars, two months before the dollar was formally devalued against gold and therefore the SDR.[3]

Interestingly, the Articles say very little about the process of quota determination. Initial quotas of original members of the Fund (those that participated in the Bretton Woods Conference and joined before December 31, 1945) are to be as agreed at Bretton Woods and are appended to the Articles: the Fund is to determine the quotas of new members. Regarding quota adjustments, the original Articles established a 'quinquennial review' of quotas; the first amendment (effective July 28, 1969) mandated a 'general review' at intervals of not more than five years. The IMF may or may not decide to propose changes in quotas as a result of these reviews. Individual members are permitted to request adjustments in their quotas at any time (so-called *ad hoc* increases in quotas), and the Fund may take such action as it sees fit. In practice, given (i) the requirement of member consent to a change in quota and (ii) the normally positive relationship of a country's benefit from membership with the level of quota, quotas are adjusted only upward.[4]

Nothing else about quota determination is laid down in the Articles. In particular, the method of determining quota levels is left open. How does the Fund allocate quotas absent constitutional guidance?

II. How the IMF actually establishes quotas

II.A. The criterion: economic size

The general principle of quota determination that has been applied both at the Bretton Woods Conference and throughout the Fund's history is that *a country's quota should be positively related to its economic size*.[5] This criterion has three corollaries. First, *given two members of unequal economic size, the larger one should have the higher quota*. This rule underlies *ad hoc* increases in quotas and increases differentiated by country or for specific countries only (what the Fund calls 'special increases') at general reviews. Second, *a new member's quota must 'fit into the existing structure of quotas*,' meaning that *its quota should be comparable to the quotas of members of comparable economic size*.

Third, *as Fund members as a group grow in economic size, their quotas should increase commensurately*, which suggests that quotas of members be increased across-the-board by a uniform percentage (what the Fund calls 'general increases' in quotas). This is a way of increasing quotas, and thereby providing the Fund with additional resources and its members with greater drawing power, without altering the relative voting strengths of economically large members.

So it is not surprising that a uniform percentage increase in quotas is often the expedient thing to do at general reviews. Indeed, of the nine general reviews and one special review thus far completed in the Fund's history, three involved zero increases in quotas and five had 'general' increases dominating 'special' increases. In the remaining two reviews, special increases were dominant; but in one of them (the eighth general review), 40 per cent of the total increase in quotas were distributed to members in proportion to existing quotas (that is, as a general increase). Further, the one-to-one switch from the dollar to the SDR as the unit of account in 1972 can be viewed as an 8.571 per cent general increase in quotas.

II.B. Quota formulas

To measure economic size, the Bretton Woods Conference made use of a mathematical formula developed in 1943 by Raymond Mikesell at the US Treasury. This so-called Bretton Woods formula was then utilized by the IMF itself, until replaced jointly by a revised Bretton Woods formula and four additional formulas all devised by the Fund staff in 1962–63. The four latter formulas were modified in 1982, while the revised Bretton Woods formula was retained. Though descriptions of these ten formulas exist, they are in scattered publications.[6] Therefore the formulas have not received the systematic and generalized treatment that follows.

The general form of the Fund's quota formulas is:

$$Q_{ijt} = (a_j\, Y_{ij} + b_j\, X_{ij} + c_j\, M_{ij} + d_j V_{ij} + e_j R_{ij})[1 + f_j(X_{ij}/Y_{ij})], \tag{1}$$

where Q_{ijt} is the calculated quota of country i at time t according to formula J; Y_{ij} is the income, X_{ij} exports, M_{ij} imports, V_{ij} variability of exports, and R_{ij} official reserves, of country i, all measured with the coverage and time span set by formula j; a_j, b_j, c_j, d_j, and e_j are non-negative coefficients fixed for j; and f_j is a parameter with value either zero or unity, as specified by formula j.

Thus the formulas embody a linear combination of measures of economic size (income, reserves, trade) with no constant term and a potential multiplicative factor that divides the formulas into a nonlinear ($f_j = 1$) and a linear ($f_j = 0$) group, depending on whether or not the factor is present. The Bretton Woods formulas are nonlinear and distinguished by $b_j = 0$, with (a_j, c_j, e_j) of the revised version half of the original. (Hence the term 'reduced Bretton Woods formula' is also used for the revised version.) Each (1962–63, 1982) pair of the four additional formulas has the same (a_j, c_j) coefficients. Two of the additional formulas are linear and two nonlinear. The 1982 nonlinear formulas have the same form as the Bretton Woods formulas ($b_j = 0$), while the 1962–63 versions involve $b_j = e_j = 0$; In the 1962–63 linear formulas, $e_j = 0$;in the 1982 modifications. all coefficients are positive and $b_j = c_j$.[7]

While the formulas are differentiated mathematically by the values of the coefficients (a_j, b_j, c_j, d_j,e_j) and the parameter f_j, they are grouped by the time period (Bretton Woods Conference, 1962–63, 1982) via the coverage and time span of the variables (Y_{ij}, X_{ij}, M_{ij}, V_{ij}, R_{ij}) that are deemed components of a country's economic size. The (original) Bretton Woods formula measured Y_{ij} as national income in 1940, M_{ij} as average annual merchandise imports over 1934–38, V_{ij} as the difference between maximum and minimum annual merchandise exports in 1934–38, and R_{ij} as gold and dollar reserves on July 1, 1943 (with all variables expressed in US dollars). The 1940, 1934–38, and 1943 dates were fixed, applying no matter what the time point (t) for which the quota (Q_{ijt}) was to be calculated.

The 1962–63 formulas involved four changes in variable measurement. First, and most importantly, up-to-date rather than historical data were applied. Second, variability of exports was redefined as one standard deviation about a moving five-year average of exports over a 13-year period. Third, each formula gave rise to two quota computations, depending on the data underlying the trade variables (X_{ij}, M_{ij}, V_{ij}): the traditional merchandise account (goods only) or the entire current account (goods, services, and private transfers). Fourth, reserves were defined as gold plus all foreign exchange rather than gold plus the dollar component of foreign exchange. Also, end-of-year reserves were taken.

The 1982 modifications widened the scope of all variables. National income was replaced by gross domestic product; current-account data were used exclusively (merchandise-trade data being dropped); and gold plus foreign exchange was replaced by total official reserves (incorporating also SDRs, European Currency Units, and reserve position in the Fund). Also, rather than end-of-year reserves, a monthly average over a year was taken.

II.C. Calculated quotas

The quota formulas were used by the IMF to determine hypothetical quotas, which came to be known as 'calculated quotas' in the Fund's terminology. At the Bretton Woods Conference and in the IMF itself from its inception in 1946 through the early 1960s, calculated quotas were obtained directly from the Bretton Woods formula. Because of the dated and fixed nature of the variables inputted into the formula, only initial quotas (whether of original or later members – and not quota increases – could be computed.

From the mid-1960s to the early 1980s, the 1962–63 formulas were used to obtain a calculated-quota range as well as a unique calculated quota, and for existing members (reflecting hypothetical quota increases) as well as new members (their hypothetical initial quotas). With five formulas and two data sets (merchandise versus current-account export and import variables), ten quota computations were made. The average of the two computations using the revised Bretton Woods formula served as one end of the range. The two lowest calculations of the four additional formulas using the merchandise-account variables, and the two lowest using the current-account variables, were together averaged to obtain the other end. The higher of the end-points constituted the unique calculated quota. In effect, this computational procedure involved a double weight for the merchandise compared to the non-merchandise components of the current account.

From the early 1980s, calculated quotas were based on the revised Bretton Woods formula and the four additional formulas as modified in 1982. The procedure was simplified, in part because only current-account data were used (thereby eliminating double weighting of the merchandise account). First (and done only for a general review), quotas computed using the four modified formulas were adjusted uniformly so that total Fund quotas equalled the sum obtained from the revised Bretton Woods formula. Second, the calculated quota was the higher of the Bretton Woods computation or the average of the two lowest figures from the modified formulas.

II.D. Actual versus calculated quotas

One might expect that actual quotas would always equal calculated quotas. However, this expectation would be far removed from reality. The *only* definitive cases in which quotas are known to have been set equal to calculated quotas are (1) three small countries at the Bretton Woods Conference, (2) four new members admitted in 1946, (3) the eighth general review of quotas in 1983, in which 60 per cent of the total increase in quotas were allocated to members in proportion to their calculated quotas, and (4) the ninth general review of quotas in 1990, in which 40 per cent of the increase were so allocated.

The main reason for the deviations of actual from calculated quotas is that measured economic size is not necessarily proportional to political strength, and quotas are ultimately a matter of political strength and negotiating skill. Therefore calculated quotas are generally used only as a basis for discussion. This was the case for the quotas generated at Bretton Woods.[8] It is also true for *ad hoc* quota increases and new-members' quotas.[9] However, the importance of political negotiation and compromise is most apparent at general reviews, as the Fund's 'inside histories' amply document.[10] The positions of countries and country groups at these reviews can be so divergent that many months are needed to reach agreement. While the 1958–59 special review (that substituted for the third quinquennial review) and the fourth and fifth reviews each took less than a year to negotiate, the sixth, seventh, and eighth general reviews required close to two years. Negotiations for the ninth general review, completed in May 1990, spanned almost three years.

III. How the IMF ought to determine quotas

III.A. Norms for quota distribution

Thus far the focus has been positive analysis, how the IMF actually determines quotas. However, the objective is to evaluate the DC versus LDC distribution of quotas. Therefore the shift must be made to normative analysis, how quotas ought to be determined. Norms for quotas can be discussed only in context of the purposes of quotas. Quotas have four main purposes (‘P’ for purpose):

P1: to determine the subscription payment;[11]
P2: to determine limits to drawings;[12]
P3: to determine allocation of SDRs;[13]
P4: to determine voting rights.[14]

These quota objectives need to be mapped onto the ultimate criteria, or norms, for quota determination, thereby justifying the norms. Such norms are, of course, much more subjective than the quota purposes. To enhance objectivity, the proposed economic norms (N1 and N2, ‘N’ for norm) are from the theory of taxation equity and are used in other international organizations to allocate country assessments. In addition, two noneconomic norms (N3 and N4) established in political science are applied:

N1: capacity to pay;
N2: benefit principle;
N3: sovereign equality;
N4: power.

Capacity to pay (N1) serves as the basis for allocating country contributions in most international organizations, notably the United Nations (UN).[15] Subscription payments to the IMF are analogous to assessed contributions to the UN, and so are appropriately based on ability to pay. A country with a greater ability to make the subscription payment should have a higher quota. Therefore the subscription-payment (P1) objective is mapped onto the capacity-to-pay (N1) norm.

The benefit principle (N2) serves as the basis of country contributions to some international organizations.[16] In the present context the principle asserts that IMF quotas should be in proportion to the benefits derived from Fund membership. The problem is that, positively, causality is reversed: benefits vary with quota level. Consider a change in interpretation so that the benefit is the potential use of, or need for, drawings and SDRs. A country’s quota level should be in proportion to its need for these facilities. Then the quota objectives pertaining to drawings (P2) and SDR allocation (P3) are mapped onto the benefit-principle (N2) norm.

Alternatively, the benefit may be that derived from the country's international transactions, which are fostered by a smoothly functioning international monetary system under the auspices of the IMF. Then the subscription-payments (P1) objective is mapped onto the benefit principle (N2). A country enjoying greater benefit from international transactions should be assessed a higher subscription payment, whence a larger quota.

Sovereign equality (N3) is a concept that involves equal treatment of countries irrespective of differential economic or power-political characteristics. Rigidly interpreted, this norm implies identical quotas and equal voting rights for all IMF members. In contrast, the power (N4) norm requires that quotas and voting rights vary with the country's economic and political power. These appear to be competing norms to which the voting-rights (P4) purpose may alternatively be mapped. However, these norms may be complementary. If sovereign equality is interpreted to mean a minimum or basic quota for each member (certainly a more realistic situation than equal quotas), a residual quota can be determined by power. In fact, all four norms can be viewed as complementary.

III.B. Economic size as a normatice criterion

The Fund's objective criterion of economic size as the determinant of quotas is readily taken also as the normative criterion. Economic size can be justified in terms of the quota objectives and norms for quota determination. An economically larger country has (i) a greater capacity to pay (P1 to N1), (ii) presumably a greater need for drawings and SDRs (P2 and P3 to N2), because the potential level of balance-of-payments disequilibria is greater, (iii) a larger volume of international transactions (P1 to N2), and (iv) greater economic and political power (P4 to N4).

III.C. Components of economic size

The issue for normative analysis of quota distribution is not the Fund's criterion – economic size – which is readily acceptable, but rather its representation in terms of the component variables employed by the IMF in its quota formulas. Two questions present themselves.

1. Are the Fund's variables justifiable in terms of the adopted norms?
2. What is the appropriate measurement of the subset of variables that can be so justified?

The five variables that enter the Fund's formulas are examined in turn.

Income is the most obvious and generally accepted indicator of economic size. There is a large and growing literature on income comparisons between countries.[17] As an indicator of ability to pay and of economic power, income is justified by the P1 to N1 and P4 to N4 mappings, respectively. On both grounds, a higher income should involve a higher quota. If a unidimensional measure of economic size were desired, income would be the clear choice.

Gross domestic product (GDP), value added within the boundary of the country, is the best measure of output for international comparison. Until 1982, the IMF used national income, an inferior measure because it involves deviations from market prices and estimates of the consumption of fixed capital.

A much more serious error is the Fund's use of the exchange rate (ER) rather than purchasing power parity (PPP) to convert income from domestic currency to dollars. *Real*, not nominal, output is the appropriate measure of economic size for international comparison; and this implies that the conversion factor reflect relative prices among the various countries.[18] Further, the PPP/ER ratio (both numerator and denominator defined as the number of units of domestic currency per dollar) *both* differs significantly from unity *and* increases with per capita income.[19] This relationship implies that converting income to a common currency via exchange rates *biases downward* the dollar output of low per-capita income countries, that is, LDCs. Therefore the Fund's calculated quotas, and thence the actual quotas, of less-developed members are unjustifiably low.[20] In sum, the income variable is appropriately measured as GDP converted to a common currency via PPP.

Reserves measure the ability of the country to make the reserve-asset portion of subscription payments. Reserves also can be viewed as an indicator of international financial strength, complementing income in representing the country's economic power. Larger official reserves therefore mean a higher capacity to pay and greater power, implying a higher quota through the P1 to N1 and P4 to N4 mappings, respectively.

However, the problem with reserves as a variable is that it can cut the level of quota the other way as well. *Lower* reserves mean a greater need, other things being equal, for drawings from the Fund and for SDRs, and therefore indicate a higher quota via the mappings of P2 and P3 to N2. In attaching a positive coefficient to reserves in its quota formulas, the IMF implicitly focuses on the P1 to N1 and P4 to N4 mappings. The P2 to N2 and P3 to N2 mappings would suggest a negative coefficient; paradoxically, a lower economic size (in its reserves component) then implies a higher quota.[21] Under either set of mappings, the reserves variable should incorporate all components of official reserves, whereas the Fund's variable (in its formulas) does so only since 1982.

Trade is a component of economic size complementing income by representing the country's importance in the international economy. Higher trade is suggestive of both larger potential balance-of-payments disequilibria and a greater volume of international transactions. Therefore a higher quota is indicated under the benefit principle by all mappings (P1, P2, and P3 to N2).

How is trade to be measured? To the mid-1960s, the Fund restricted trade to the merchandise account; thereafter, all of the current account (goods, services, and private transfers) were incorporated. Transfers are not part of output, and also it is illogical to exclude official but not private transfers; so the trade variable here pertains to the goods-and-services account. Defining the variable straightforwardly as the sum of receipts and payments is most logical.

The IMF unnecessarily complicates matters by introducing several asymmetries. In some formulas imports and exports have different coefficients. There is no good reason to weight them differently. In all formulas the variability of exports is included as a variable, either in place of the level of exports or in addition to it. The asymmetry of combining the *level* of imports with the *variability* of exports in a formula has no justification. Symmetry could be reached by incorporating the variability of imports as an additional variable; but then the two variables would offset each other and one would better measure the variance of the overall balance of payments directly, which brings one far from the direct criterion of economic size. Also, it is not clear that past payments disequilibria are an appropriate indicator of future deficits. Finally, the export/income ratio, included as a multiplicative factor in some formulas, is not only asymmetrical in excluding imports but also introduces nonlinearities that can lead to strange and unintended implications for quota levels.[22]

IV. Methodology for evaluating quotas

A 'quota disparity' is defined as the deviation of an actual quota from the corresponding quota that would be predicted from appropriately measured economic size, that is, from the income, reserves, and trade variables adopted in Section III. If the actual quota exceeds the predicted quota, the country is favored in the distribution of quotas; if the actual is below the predicted quota, the country is in a disadvantaged position. The objective is to examine if quota disparities have historically been to the disadvantage of LDCs and whether the position of this country group has worsened or improved over time.

IV.A. Time and unit dimensions of variables

Quota distributions at various end-of-year points over the Fund's history will be selected for examination. The variables that are used to predict the normative quota are logically defined contemporaneously. If actual quotas are for the end of year t, then reserves are also end-of-year t, while the flow variables (income and trade) are for the entirety of year t.[23] For consistency, quotas and the evaluative variables are all converted to US dollars – quotas, reserves, and trade by exchange rates, income by purchasing power parties.[24]

IV.B. Dates for quota examination

From the Fund's inception to the end of 1989, general reviews accounted for 83 per cent of total Fund quotas, membership (initial quotas) 13 per cent, *ad hoc* increases 3 per cent, and special policies (see note 29) 1 per cent. Therefore, in selecting the dates at which to evaluate the quota distribution, general reviews should be the governing factor: it is important to avoid selecting a date over which subscription payments for the quota increases emanating from a review are in process. Further, end-of-year quotas and five-year reference dates (following the Fund's policy of no more than five years between general reviews) are obvious points of retrospection.

The dates (year-ends) selected are 1950, 1960, 1965, 1971, 1975, 1981, and 1985. The year 1971 rather than 1970 is taken to allow completion of payments for quota increases under the fifth general review. The year 1981 is chosen over 1980 to encompass an *ad hoc* increase in Saudi Arabia's quota, which was the biggest such increase in the Fund's history to the end of 1989 (constituting 38 per cent of the sum of all *ad hoc* increases over 1946–89).

The beginning dates, 1950 and 1955, involve membership as virtually the sole source of quotas (in fact, accounting for 99 per cent of total Fund quotas at the end of 1955), and so these dates are innocuous. The early quinquennial reviews did not result in quota increases, and special policies to adjust quotas were not yet in existence. The final date, 1985, incorporates full subscription payments for quota increases emanating from the eighth general review. There were no quota changes at all in 1986–89, and initial quotas of only three new members.

IV.C. Country sample

For any year, the country sample is the set of Fund members for which data exist to construct *all* the evaluative variables: income, trade, and reserves. In addition, China is excluded from the sample in 1950–75. Until 1980, China was represented in the Fund by the Republic of China, the government of which had been ousted from the Chinese mainland by April 1950. However, China was an original member of the Fund, and its quota (determined at Bretton Woods) assumed sovereignty over the mainland. Therefore China's initial quota was viewed, quite correctly, as far greater than that justified by the economic size of the territory (principally Taiwan) controlled by the Republic of China, and China received no quota increase until after its representation was switched to the People's Republic of China in April 1980. Inclusion of China in the 1950–75 samples would mean an outlier observation, and of an extreme magnitude in the early years (before Taiwan experienced substantial economic growth). In 1981 and 1985, with the represented government in control of the mainland, China is legitimately included in the sample.

As shown in Section II of Table 6.1, the size of the sample (number of countries) increased by 150 per cent from 1950 to 1985, but the percentage of Fund membership incorporated in the sample decreased. However, the sample retained – in fact, gained – richness in terms of the importance of the included countries. The sample share of total Fund quotas increased

Table 6.1 LDCs in the Fund

Item	*1950*	*1955*	*1960*	*1965*	*1971*	*1975*	*1981*	*1985*
I. Total Fund membership								
Number of countries[a]	49	58	69	103	120	128	143	149
LDC percentage: countries[a]	61	64	67	76	78	79	81	81
LDC percentage: quotas[a]	21	21	24	27	27	28	35	33
II. Sample of Fund members: properties								
Number of countries[a]	45	55	64	92	103	110	115	114
Percentage of countries[a]	92	95	93	89	86	86	80	77
Percentage of quotas[a]	91	93	96	96	97	97	98	97
III. Sample of Fund members: LDC percentages								
Countries[a]	60	62	64	73	75	75	77	75
Quotas[a]	15	15	20	24	25	26	33	31
Income	20	20	23	24	26	28	39	41
Trade	19	16	17	17	16	21	24	20
Reserves[a]	16	15	13	15	16	29	25	26

a End-of-year.

from 91–93 per cent in 1950–55 to 96–98 per cent in 1960–85. Countries excluded from the sample tended to be smaller countries after 1955.[25]

IV.D. Country classification

Members are grouped into the developed and less-developed categories, following closely the Fund's own classification scheme presented in its publications.[26] For some of the analysis, it is useful to divide the developed members into 'industrial countries' and 'other developed countries.' The grouping of industrial countries expands from 14 to 20 members for 1981 and 1985, according to the Fund's classification and honored here.[27] 'Other developed countries' comprises eight members in 1985.[28]

V. Aggregate analysis

In aggregate analysis, the total quota of LDCs is evaluated relative to the total quota of DCs. The less-developed and the developed members of the Fund are each considered a 'supercountry,' with the distribution of quotas within each supercountry irrelevant.

Consider first total Fund membership. Section I of Table 6.1 makes the important point that *LDCs have low quotas*. Consistently over 1950–85, the LDC share of Fund membership runs 40–50 percentage points higher than its share of quotas. Turning to the sample of Fund members (Section III of Table 6.1), the tendency of LDCs to have low quotas (LDC percentage of countries exceeding LDC percentage of quotas) repeats itself. In addition, the normative justification is clear: LDCs *as a group* are economically small. Throughout 1950–85, the LDC percentage of countries is far greater than its percentage of income, trade, or reserves. At first glance, on an aggregate basis, the low quotas of LDCs simply reflect their small economic size.

However, the actual numbers tell a more complicated story. LDC percentages of income are uniformly greater than their percentages of trade and reserves (except for the anomalous year 1975, associated with a tremendous shift of reserves in favour of oil-exporting countries). This relationship suggests that, in evaluating quotas by their correspondence to economic size, measuring size by income is favorable and by trade or reserves unfavorable to LDCs relative to DCs.

In fact, the LDC share of quotas over 1950–85 is always less than (or equal to) that justified by its share of income, and often more than that justified by its shares of trade and reserves. If real income is the best measure of economic size, then, since 1950, LDCs suffered an unfavorable quota disparity (the difference between the LDC percentage of quotas and of income), which declined to nil in 1965, but subsequently increased steadily to 1985. The disparity was greater in 1981 and (especially) 1985 than it had ever been. On the other hand, a favorable quota disparity for LDCs (LDC percentage of quotas exceeding LDC percentage of evaluative variable) was the general rule for trade and reserves over 1950–85 and was greatest at the end of the period for trade, in the middle of the period for reserves.

VI. Disaggregate analysis

In disaggregate analysis, the units of observation are individual countries, but the focal point remains differential treatment of LDCs versus DCs as groups. Two techniques are used:

1. Ratio comparisons, in which income is viewed as the primary measure of economic size, and quota/income and related ratios are examined.
2. Regression analysis, which involves evaluation of quotas using income, trade, and reserves simultaneously, and also allows for country-group effects independent of measures of size.

VI.A. Ratio comparisons

If income is considered the best unidimensional measure of economic size, then the quota/income (Q/Y) ratio becomes the variable of interest. Define the 'normalized' quota/income ratio for a given country in a given year as its Q/Y value divided by the sample mean (Q/Y). The expected value of a country's normalized quota/income ratio is unity. A ratio greater than unity constitutes a favorable quota disparity, that is, a quota greater than that justified by economic size relative to the average treatment of all countries. Similarly, a ratio less than unity represents an unfavorable quota disparity.

Results are presented in Table 6.2 in terms of the percentage of DCs or LDCs with favorable quota disparities. A percentage above 50 denotes a favorable disparity, below 50 an unfavorable one. There was a precipitious decline until 1971 in the initially highly favorable quota disparities accorded DCs and little change since then. In contrast, only 22 per cent of LDCs enjoyed a favorable quota disparity in 1950; the figure increased substantially to 1965 and then fell slowly to 1985. Through 1960, the percentage of favored DCs exceeded that of LDCs; but by 1965 the percentage of favored LDCs exceeded that of DCs and continued to do so to 1985.

Continuing the ratio analysis with additional evaluative variables, the theory adopted is that income is the most important normative determinant of quota, but the quota/income (Q/Y) ratio is justifiably pushed up by a high trade/income (T/Y) or reserves/income (R/Y) ratio:

$$Q/Y = F(T/Y, R/Y), \qquad (2)$$

where $F' > 0$ for all arguments.

A country's normalized quota/trade ratio is its normalized quota/income ratio *divided by* the ratio of its T/Y value to the sample mean (T/Y), and similarly for trade (T) replaced by reserves (R). Again the expected value of the normalized ratios is unity. Therefore, if the normalized quota/trade and quota/reserves ratios are both less than unity, then the quota/income ratio, and therefore the country's quota itself, is unambiguously lower than the level justified by trade and reserves together, in relation to income, implying

Table 6.2 Normalized ratios of quota to explanatory variables

	Percentage of countries with			
	Quota/income[a] *greater than unity*		*Quota/trade, quota/reserves*[b] *greater than over less than unity*[c]	
Year[d]	*DCs*	*LDCs*	*DCs*	*LDCs*
1950	78	22	50	–15
1955	67	26	33	–6
1960	48	34	–13	22
1965	32	46	–60	29
1971	23	42	–54	42
1975	26	41	–37	34
1981	22	40	–49	33
1985	25	38	–67	30

a A country's normalized quota/income ratio is its quota/income value divided by mean (quota/income), where the mean is taken over all countries in the sample.
b A country's normalized quota/trade ratio is its normalized quota/income ratio divided by its normalized trade/income ratio, where the latter ratio is the country's trade/income value divided by mean (trade/income). An analogous definition applies to the normalized quota/reserves ratio.
c Excess of the two ratios greater than unity over the two ratios less than unity.
d End-of-year for quota and reserves.

an unfavorable quota disparity. The opposite is true for the ratios exceeding unity.

If the percentage of countries with normalized quota/trade and quota/reserves ratio both above unity exceeds the percentage with both ratios below unity, the country group may be said to have a favorable quota disparity. By this criterion, LDCs had an adverse quota disparity only in 1950 and 1955, and a favorable one thereafter; the opposite is true for DCs (see Table 6.2). The positive quota disparity of LDCs peaks in 1971; the negative disparity of DCs reaches its maximum in 1985.

VI.B. Regression analysis

Consider the following equation to evaluate the distribution of quotas at a given date (specific year-end):

$$Q_i = a + bY_i + cT_i + dR_i + eP_i + fIND_i + gOTH_i + u_i, \tag{3}$$

where Q_i denotes the quota of country *i*, Y_i real income, T_i trade, R_i reserves (all in millions of dollars), and P_i population (millions of persons). The variables IND_i and OTH_i are dichotomous variables, unity if country *i* fits the industrial or other-developed category, respectively, 0 otherwise. Parameters of the equation are *a* to *g*, and u_i is the stochastic disturbance for country *i*.

This equation combines variables that measure economic size for normative analysis with variables representing favoritism. The income, trade, and reserves variables together represent economic size. It is expected that their parameters (*b, c, d*) are positive; a country of greater economic size in any dimension, with other dimensions of size held constant, should have a higher quota.

The constant term, *a*, is 'autonomous quota,' the amount of quota accruing to every country independent of its economic size. Determinants of autonomous quota include the Fund's formal policies on the matter – its small-quota policy, compensatory financing facility, and differential rounding-up of uniform percentage increases in quotas – but are not limited to these policies; for autonomous quotas are the result of many Fund actions on quotas.[29] A positive constant respects the sovereign equality of countries and finds justification in the P4 to N3 mapping (see Section III.A).

The dichotomous variables, IND_i and OTH_i, seek to measure 'autonomous favoritism' toward one or the other group of DCs, that is, favoritism not related to economic size. Positive values of *f* and *g*, the coefficients of these variables, imply a systematic quota disparity in favor of DCs to the detriment of LDCs.

Population is a continuous variable that also denotes favoritism, but toward LDCs. The LDC share of population (61–69 per cent in 1950–75 and 79 per cent in 1981–85) is about the same as the LDC share of countries in the IMF (see Section III of Table 6.1). If population were the criterion of economic size (which it decidedly is not), LDCs would no longer be small relative to DCs. So a positive value of the parameter e denotes a quota disparity in favor of LDCs over DCs. Another way to see this is to deflate all variables by income, as in equation (2). Then the quota/income (Q/Y) ratio depends on population/income (P/Y), among other variables. Per capita income, generally low in LDCs, is the inverse of P/Y. A positive coefficient on P/Y in the transformed extended equation (2) or, correspondingly, on P in the original equation (3), rewards lower per capita income with a higher quota.

The remaining term in equation (3) is u_i, the stochastic disturbance for country i in the given year. To estimate equation (3), assumptions must be made regarding the disturbance.[30] It is reasonable to assume that u_i is normally and independently distributed (with mean zero) for all i; but decidedly *unreasonable* to impose an identical distribution, because the variance of the disturbance term for the quota – just as the quota itself – is expected to be positively related to economic size. With real income the primary measure of economic size, the common heteroscedastic specification may be made:

$$\text{variance}\ (u_i) = kY_i^2 \tag{4}$$

where k is a positive parameter.

The generalized least-squares estimates of equation (3) are presented in Table 6.3 for the eight sample years (identified in the first column).[31] Regression coefficients (with t-values in parentheses) are shown in columns beneath the respective variables. The measures of goodness of fit, $\bar{R}^2$ and r^2, are as defined in note 31. Variables making a negative contribution to $\bar{R}^2$ are dropped from the regression, and the equation is re-estimated.

Economic size is measured as $(bY_i + cT_i + dR_i)$. The coefficient, b, on income is positive and significant (meaning significantly different from zero) in all years except 1965. Recalling that income (unlike trade and reserves) is a variable that enhances the economic size of LDCs relative to DCs, the larger coefficients in 1975–85 compared to earlier years suggest greater favoritism toward LDCs in the measurement of economic size toward the end of the 1950–85 period. The coefficient on trade is significantly positive in all years except 1955. The lower coefficients from 1975 onward mean reduced favoritism toward DCs. Similarly, the negative coefficient on reserves in 1971–75 followed by nonsignificance in 1981–85 implies favoritism for LDCs. The nonsignificance or negative sign of the coefficient on reserves in all but one year perhaps reflects reserves as an indicator of need for drawings and SDRs offsetting reserves as a measure of economic size (see Section III.C). The strong conclusion is that *the coefficients on all three components of economic size exhibit greater favoritism toward LDCs in the quota distribution in the latter part of the 1950–85 period.*

The constant term measures that part of quota equally distributed to all countries and therefore independent of size. Because LDCs tend to be economically small, a large constant is in their interest. Therefore the negative

Table 6.3 Regression equations for quota

Year	*Income*	*Trade*	*Reserves*	*Population*	*Constant*	*IND*	*OTH*	$\bar{R}^2$	r^2	*df*
1950	0.0020**	0.0222**			–3.09**	5.33*	1.77	0.80	0.98	40
	(4.25)	(3.74)			(3.77)	(2.16)	(1.13)			
1955	0.0013**	0.0068	0.0521**		–1.61**	7.48**		0.80	0.93	50
	(3.88)	(1.83)	(3.16)		(2.68)	(3.64)				
1960	0.0022**	0.0114*	0.0383		1.18		5.37**	0.75	0.94	59
	(4.76)	(2.15)	(1.72)		(0.90)		(2.90)			
1965		0.0194**	0.0223	1.20**	4.02**			0.72	0.87	88
		(6.84)	(1.38)	(5.94)	(5.89)					
1971	0.0018**	0.0224**	–0.0239	1.34**	6.25**	–5.26	4.53	0.78	0.95	96
	(2.94)	(6.28)	(1.87)	(3.84)	(9.31)	(1.85)	(1.91)			
1975	0.0029**	0.0077**	–0.0158*	1.38**	7.60**	–12.92	8.93*	0.66	0.93	103
	(3.61)	(4.30)	(2.16)	(2.90)	(8.23)	(1.05)	(2.29)			
1981	0.0054**	0.0037**		2.94**	14.68**			0.53	0.78	111
	(3.57)	(2.79)		(3.15)	(6.30)					
1985	0.0046*	0.0098**		3.82**	17.69**			0.55	0.76	110
	(2.22)	(4.04)		(3.27)	(5.67)					

Notes:
IND = 1 if industrial country. 0 otherwise.
OTH = 1 if other developed country, 0 otherwise.
$\bar{R}^2$ = R^2 corrected for degrees of freedom. Special definition of R^2 is used: see note 31.
r^2 = square of the correlation coefficient between actual and predicted quota.
df = degrees of freedom.
* Significant at 5 per cent level.
** Significant at 1 per cent level.

coefficient in 1950–55 is unfavorable, the positive constant thereafter favorable toward LDCs. *The steadily increasing absolute value of the constant over 1950–85 involves increasing favoritism toward LDCs*. The Fund's efforts, begun in 1956, to provide members with a component of quota not governed by economic size have had significant and ongoing effect, that survived even the termination of the 'small-quota polity' and the related policy under the compensatory financing facility.

Population exhibits a coefficient that moves from nonsignificance in 1950–60 to a steadily increasing (positive) value highly significant throughout 1965–85, involving ever-greater favoritism for LDCs. In contrast, the dichotomous variables (*IND* and *OTH*) representing favoritism for DCs have only isolated significantly positive coefficients that cease by 1981.

The explanatory power of the regression deserves comment. Both $\bar{R}^2$ and r^2 are lowest in 1981–85. The larger, non-explainable element of quotas at the end of the 1950–85 period may reflect the strong 'inertia' of relative quotas even in the face of changing relative economic sizes of Fund members. There is a strong tendency at general reviews to have an across-the-board uniform percentage increase in quotas (see Section II.A). While *ad hoc* quota increases inherently alter relative quotas, as did the now-defunct small-quota policy and the quota aspect of the compensatory financing facility, their quantitative impact is small. From the Fund's inception to the end of 1989, *ad hoc* increases and these special policies accounted for less than 4 per cent of total quotas.

It is exceedingly important, therefore, for new members concerned with their ongoing relative quota to obtain as high an initial quota as possible.

VII. Principal conclusions

1. The Fund has always utilized mathematical formulas to calculate quotas for existing and prospective members from measures of their economic size. However, there is good evidence that actual quotas can deviate substantially from corresponding calculated quotas.
2. The IMF criterion of economic size for quota distribution receives normative justification in terms of quota objectives (the determination of subscription payments, drawings limits, SDR allocation, and voting rights) and the ultimate norms of quota distribution (capacity to pay, benefit principle, power).
3. 'Autonomous' quota, a component of quota independent of economic size, receives normative justification in terms of the sovereign equality of states. A large autonomous quota is to the advantage of LDCs because of their small economic size.
4. *Appropriately measured* income, reserves, and trade receive normative justification as representations of economic size. A favorable quota disparity is defined as an actual quota in excess of the quota predicted by income, reserves, and trade – the components of economic size.

5. The Fund's measurement of the income component of economic size is both methodologically deficient and detrimental to LDCs, because exchange rates rather than purchasing power parities are used to convert data to a common currency.
6. Income (based on purchasing power parities) is a variable that enhances the economic size of LDCs relative to DCs and therefore should act to increase LDC versus DC quotas, whereas the opposite is true of trade and reserves.
7. For the 1950–85 period, there are four main empirical findings: (a) The quota disparities in favor of LDCs and those against them are about evenly balanced. (b) In the main, favoritism toward LDCs increases over time, though sometimes stabilization occurs in mid-period. (c) Toward the end of the period, such favoritism is at a peak more often than the reverse. (d) There is a strong inertia of relative quotas.

Appendix: the data

Quota End-of-year figures from IMF, *International Financial Statistics (IFS)*, various issues. The 1975–85 data are converted to dollars using the end-of-year dollar/SDR exchange rate.

Income and Population Data on real GDP per capita and on population (with income the product of the two) are taken principally from Summers and Heston (1988). Using an overlapping year to make the data consistent, missing observations are obtained from Summers *et al.* (1980), Kravis *et al.* (1982), Summers and Heston (1984), and *IFS, 1987 Yearbook*.

Trade The sum of gross credits and debits on goods-and-services account, in US dollars, for the calendar year (except where only another annual basis is available; and where that basis is July 1 to June 30, the average of the two figures overlapping the calendar year is taken). The primary source is IMF, *Balance of Payments Statistics*, especially the Yearbooks.

For 1950 and 1955, net components of UK services are converted to gross format using 1957 data as an overlap. For all years, trade (available only in consolidated form) is allocated between Belgium and Luxembourg in proportion to their relative incomes.

Where balance-of-payments data are unavailable, national-accounts export and import (goods-and-services) data are used to estimate the missing observations on the basis of overlapping figures for an adjacent year. (For Lesotho, 'net factor income from abroad' is added to national-accounts exports, because of the dominance of labor income in balance-of-payments exports.) Where national-accounts data are unavailable or inconsistent, resort is had to merchandise export and import figures.

Reserves Defined as the sum of gold, foreign exchange, SDRs, and reserve position in the Fund, expressed in millions of dollars. The source is *IFS*, various issues. Gold is valued at the official price in 1950–71 and the London market price in 1975–85. Nongold reserves in 1971–85 are converted from SDRs to dollars using the end-of-year SDR/dollar exchange rate.

Notes

* Reprinted from *Journal of International Money and Finance*, vol. 10, Lawrence H. Officer, 'Are International Monetary Fund Quotas Unfavorable to Less-Developed Countries? A Normative Historical Analysis,' pp. 193–213. Copyright 1991, with permission from Elsevier.

1 The basic sources of information about the IMF are the *Annual Report* of the Fund's Executive Board, which began in 1947, and the Fund's three inside-histories: Horsefield and de Vries (1969) and de Vries (1976, 1985).

2 The only known empirical investigation of DC versus LDC quotas is de Vries (1985, Vol. I, pp. 536–538). De Vries simply computes quota shares by country group for selected dates, without analysis or evaluation.

3 Upon its creation in 1969, the SDR was given a fixed gold value (SDR 35 = 1 oz) equal to that of the dollar ($35 = 1 oz), and therefore one SDR equalled one dollar. In December 1971, the United States committed itself to devalue the dollar to $38 an ounce. Effectively, the SDR/dollar rate immediately became SDR 1 = $1.08571 (as the Fund's exchange-rate statistics, published in its *International Financial Statistics*, were later to show); but the conversion of the Fund's accounts from dollars to SDRs in March 1972 maintained the legal equivalence SDR 1 = $1 up to the dollar's formal devaluation in May. In February 1973, when the dollar was further devalued to $42.22 an ounce, the rate became SDR 1 = $1.20635. Beginning July 1974, the SDR has been valued as a basket of currencies (originally with a sixteen-currency, since 1981 with a five-currency composition); and since that date, the SDR has fluctuated against the dollar.

4 For the principal benefits related to quota level, see notes 12–14. While there are also costs that increase with quota (such as the reserve-asset portion of subscription payment and, under certain circumstances, the domestic-currency portion, via use of the member's currency in Fund transactions), the revealed preference of countries suggests that there is typically a strong net benefit from a given quota and one that increases with the quota level. At Bretton Woods all negotiators argued for high quotas for their respective countries, and at one time ten of the 44 countries assigned quotas formally complained that their quota was too low (Horsefield and de Vries, 1969, vol. I, p. 98). Throughout the Fund's history, there have been many cases of requests for *ad hoc* quota *increases*, but only one instance of a requested quota decrease (Honduras in 1948, a temporary decrease – see Horsefield and de Vries, 1969, vol. I, pp. 196, 302). Rarely have countries refused quota increases proposed at general reviews or consented to an amount lower than the proposed increase. There were only 17 such instances from the fourth quinquennial review through the eighth general review. Some rejections were probably made on political grounds – the Republic of China, aware of its oversized quota and fearful that it might be replaced by the People's Republic, Democratic Kampuchea (Cambodia) and Iran out of antipathy toward the international organization at the time.

5 The criterion of economic size is implicit in the Fund's behavior rather than embedded as formal IMF doctrine. Rarely, IMF authors make the criterion explicit. According to Edo (1978, p. 166), quotas are supposed to 'be comparable

for economies of comparable size.' Chandavarkar (1984, pp. 11, 12) writes: 'Initially, the size of a member's quota is calculated to reflect its relative economic size in the world economy . . .' and 'Reviews and adjustments of quotas are designed to reflect the changing relative economic size and circumstances of member countries.' De Vries (1985, vol. I, p. 515) notes that, to assign the quota of a new member, 'the economic size of the country was compared with that of existing members.'

6 The Bretton Woods formula is presented incorrectly in a number of places but correctly in Altman (1956, pp. 138–139). Edo (1978, p. 166), and Hooke (1981, p. 10, n. 5). The additional 1962–63 formulas are in Edo (1978, p. 166), while the revised Bretton Woods formula and the 1982 formulas are in Chandavarkar (1984, p. 15, n. 3).

7 Exact values of the coefficients (a_j, b_j, c_j, d_j, e_j) for each formula are found among the sources listed in note 6.

8 As Altman (1956, pp. 140–141) writes, 'the formula calculations were intended to serve only as a starting point for further discussion, and quotas themselves were set by a process of negotiation and compromise.' See also Horsefield and de Vries (1969, vol. I, pp. 97–98).

9 The IMF does not present quotas to prospective members on a 'join with it or don't join basis;' and there is considerable scope for negotiation between an applicant and the Fund on the initial quota level. 'In a number of instances, negotiations have been prolonged by the applicant's wish to have a larger quota than the one calculated either on the basis of formulas or by comparison with the quotas of existing members' (Gold, 1974, p. 171).

10 See Horsefield and de Vries (1969, vol. I, p. 583) and de Vries (1976, pp. 290–297; 1985, pp. 517–535).

11 The Articles state that a member's subscription equals its quota and that the typical division of payment is 25 per cent in reserve assets and 75 per cent in the member's own currency. Members must, of course, subscribe both to their initial quota and to quota increases.

12 A member may make drawings from the IMF subject to limits defined as given percentages of quota.

13 SDRs are allocated to members strictly as a uniform percentage of quota, when this reserve asset is created by the Fund.

14 A member is assigned 250 'basic votes' (independent of quota) plus one vote for each SDR 100 000 of its quota.

15 See Pincus (1965, pp. 59–61, 95–98).

16 The Universal Postal Union, Hydrographic Bureau, General Agreement on Tariffs and Trade, and Bureau of Customs Traffic, according to Pincus (1965, pp. 59, 92–94).

17 Recent surveys are provided by Falvey and Gemmell (1989) and Officer (1989).

18 'The PPP [rather than exchange-rate] way is the right way to go, by definition, because PPP is the ratio of the domestic cost of buying a bundle of goods and services in the country at its own prices to the corresponding cost in dollars of the same bundle in the United States' (Heston and Summers, 1988, p. 467).

19 The evidence for the positive correlation of PPP/ER with per capita income is threefold. First is the 'casual empirical observation that services and nontradables generally are relatively cheap in low-income countries' (Kravis and Lipsey, 1983, p. 11). In the history of economic thought, those who made this observation include David Ricardo in 1817. Frank Taussig in 1928, Roy Harrod and Bertil Ohlin in the 1930s, Everett Hagen and Kurt Rothschild in the 1950s, and Dan Usher, Bela Balassa, and Paul Samuelson in the early 1960s. (For references, see Kravis *et al.*, 1978, p. 219; and Officer, 1982. pp. 126–127.) Second, there are a variety of economic theories (some proposed by the above writers) that predict a positive

correlation of PPP/ER with per capita income: (i) a productivity-differential model, (ii) a factor-proportions model, (iii) a specific-factors model, and (iv) an economies-of-scale model; presented well in Kravis and Lipsey (1983, pp. 4–5, 11–12), Bhagwati (1984), Clague (1985), and Panagariya (1988), respectively. Third, a group of econometric studies confirms the correlation: Clague (1986, 1988a, 1988b), Falvey and Gemmell (1989), Kravis *et al.* (1983), Kravis and Lipsey (1983, 1987, 1988), and Officer (1989).

20 As early as 1969, less-developed members of the IMF complained that the use of exchange rates to convert income to dollars yielded unfairly low quotas for LDCs. See de Vries (1976, vol. I, pp. 291, 304–305).

21 In fact, it has been argued that 'if reserves are to enter the formula at all, it would be more consistent to make quotas negatively correlated to reserves' (Pincus, 1965, p. 100).

22 See Horsefield and de Vries (1969, vol. II, p. 360), Pincus (1965, p. 100), and Edo (1978, p. 166).

23 Of course, the Fund lacks up-to-date figures for the variables entering its quota formulas; so its computations are necessarily out of date (and in the case of the original Bretton Woods formula, deliberately so). The dated quality of calculated quotas may be an explanation of the important role of negotiations in quota determination (see Section II.D).

24 Unlike income; quotas, reserves, and trade may legitimately be converted to a common currency using the current exchange rate – quotas and reserves because output is not being compared internationally, trade because a 'law of one price' for traded goods may be applied with far more justification than for all output.

25 High-quota members outside the sample (due to lack of data) are Czechoslovakia (1950). Cuba (1950–60), Libya (1960–85), Iraq (1981–85), and Iran (1985).

26 See de Vries (1976, vol. I, p. 82, n. 5; 1985, vol. I, p. 42, n. 11) and International Monetary Fund, *Annual Report* and *International Financial Statistics*, various issues.

27 The original 14 industrial countries (only 11 of which were members in 1950) are Austria, Belgium, Canada, Denmark, France, Germany, Italy, Japan, Luxembourg. Netherlands, Norway, Sweden, the United Kingdom, and the United States. The March 1980 issue of *International Financial Statistics* added Australia, Finland, Iceland, Ireland, New Zealand, and Spain to the list. Initially, only the Group of Ten (the 14 countries excluding Austria, Denmark, Luxembourg, and Norway) were classified as industrial, an unnecessary restriction for the present study.

28 Greece, Malta, Hungary, Portugal, Romania, South Africa, Turkey, and Yugoslavia. Of course, the six additions to the industrial group (see note 27) were shifted from other-developed status. It should be noted that the category 'other developed countries' was called various names through the Fund's history ('other high-income countries,' 'other developed areas,' 'primary producing countries in more developed areas,' and 'more developed primary producing countries'). In fact, this category was dropped by the IMF in 1980; but, for a consistent analysis, it is continued here in 1981 and 1985, with Hungary logically added to the group for the latter year.

29 In January 1956, the IMF initiated its small-quota policy, under which members were guaranteed a minimum quota irrespective of economic size, but the policy began to be eroded in 1961 and was abandoned in 1967. In connection with its 'compensatory financing facility,' established in 1963 and providing for drawings oriented to primary-producing countries, the Fund gave sympathetic consideration to requests for increases in quotas. All such requests for special quota increases having been granted, and the residual entitlements having been incorporated in the fifth general review of quotas in 1969, this policy of special quota increases was deemed completed and was terminated with that review.

In the fourth and fifth general reviews of quotas (in 1965 and 1970), the increase in autonomous quota was incidental; for the amount of rounding-up *increased* with quota level. In contrast, the eighth and ninth general reviews actively increased autonomous quota: quotas were rounded up to the next higher multiple of SDR 0.5 (0.1) million for new quotas at or below (above) SDR 10 million; so the rounding amount was higher for low quotas.

30 The substantive differences between equations (1) and (3) are instructive. First, equation (1) represents Fund formulas that aid in the determination of quotas, whereas equation (3) serves purely to evaluate quotas. Second, the coefficients of equation (1) are known numbers, whereas those of equation (3) must be estimated. Third, the Fund formulas, in contrast to equation (3), have no constant term and so impose zero 'autonomous quota.' Fourth, equation (1), unlike equation (3), allows for a nonlinear term.

31 It is well known that the generalized least-squares estimator of equation (3) is equivalent to the weighted (ordinary) least-squares estimator with all variables deflated by Y_i; for the transformed disturbance, u_i/Y_i, is homoscedastic with variance k for all i. Kmenta (1986, p. 283) says the advantage of equation (4) to specify heteroscedasticity is its simplicity but criticizes its rigidity. In this case, however, with real income the primary component of size, the transformed equation makes economic sense; for it relates Q_i/Y_i to its determinants in the spirit of equation (2).

As is the case with most generalized least-squares estimators, there is the problem that no one measure of 'goodness of fit' (R^2) of the regression satisfies all desirable criteria. The criteria for R^2 are that it (1) measure how well the regression equation fits the sample points, (2) equal the proportion of the variation of Q_i explained by the regression, (3) measure the equation's predictive power over the sample, (4) monotonically relate to the F-statistic used to test the hypothesis that all regression parameters (except the constant) are zero, and (5) measure Q_i in original units rather than in terms of a transformation. The squared correlation, r^2, between the actual and estimated Q_i fulfills (3) and (5) but not the other criteria. As a complementary measure, consider

$$R^2 = 1 - \Sigma w_i \hat{u}^2_i / \Sigma w_i (Q_i - \bar{Q})^2, \tag{5}$$

where summation is over i, $\hat{u}_i$ is the residual for country i, $\bar{Q}$ is the mean of Q_i weighted by w_i, and w_i is $1/Y_i^2$. This R^2 satisfies (1), (2), and (4). Between them, the two measures fulfill all five criteria and so are jointly adopted for this study. The R^2 is adjusted for degrees of freedom (and denoted as $\bar{R}^2$).

Complete presentation of the issue of goodness of fit under generalized least-squares estimation is absent in the literature. The best discussion is Judge *et al.* (1985, pp. 29–32).

References

Altman, Oscar L., 'Quotas in the International Monetary Fund,' International Monetary Fund *Staff Papers*, August 1956, **5**: 129–150.

Bhagwati, Jagdish N., 'Why are Services Cheaper in the Poor Countries?' *Economic Journal*, June 1984, **94**: 279–286.

Chandavarkar, Anand G., *The International Monetary Fund: Its Financial Organization and Activities*, Pamphlet Series No. 42, Washington, DC: International Monetary Fund, 1984.

Clague, Christopher K., 'A Model of Real National Price Levels,' *Southern Economic Journal*, April 1985, **51**: 998–1017.

Clague, Christopher K., 'Determinants of the National Price Level: Some Empirical Results.' *Review of Economics and Statistics*, May 1986, **68**: 320–323.

Clague, Christopher K., 'Explanations of National Price Levels,' in Jorge Salazar-Carrillo and D. S. Prasada Rao, eds, *World Comparison of Incomes, Prices and Product*, Amsterdam: North-Holland, 1988, 237–262 (1988a).

Clague, Christopher K., 'Purchasing-Power Parities and Exchange Rates in Latin America,' *Economic Development and Cultural Change*, April 1988, **36**: 529–541 (1988b).

De Vries, Margaret G., *The International Monetary Fund, 1966–1971: The System Under Stress*, 2 vols, Washington, DC: International Monetary Fund, 1976.

De Vries, Margaret G., *The International Monetary Fund, 1972–1978: Cooperation on Trial*, 3 vols., Washington, DC: International Monetary Fund, 1985.

Edo, Michael E., 'Multiformula Method Adds Flexibility in the Calculation of Members' Quotas,' *IMF Survey*. June 5, 1978, **7:** 166–168.

Falvey, Rodney E., and Norman Gemmell, *Explaining Service Price Differences in International Comparisons*. Research School of Pacific Studies, Australian National University, Working Papers in Trade and Development No. 89/2, May 1989.

Gold, Joseph, *Membership and Nonmembership in the International Monetary Fund*, Washington, DC: International Monetary Fund, 1974.

Heston, Alan, and Robert Summers, 'What We Have Learned about Prices and Quantities from International Comparisons: 1987,' *American Economic Review*, May 1988, **78**: 467–473.

Hooke, A. W., *The International Monetary Fund: Its Evolution, Organization, and Activities*, Washington, DC: International Monetary Fund, 1981.

Horsefield, J. Keith, and Margaret G. De Vries, *The International Monetary Fund, 1945–1965: Twenty Years of International Monetary Cooperation*, 3 vols, Washington, DC: International Monetary Fund, 1969.

International Monetary Fund, *Annual Report*, various issues.

International Monetary Fund, *Balance of Payments Statistics*, various issues.

International Monetary Fund, *International Financial Statistics*, various issues.

International Monetary Fund, 'Members' Quotas Affect Their Access to Fund Resources,' *IMF Survey*, September 1988, **17**: 4.

Judge, George G., W. E. Griffiths, R. Carter Hill, Helmut Lütkepohl, and Tsoung-Chao Lee, *The Theory and Practice of Econometrics*, 2nd edition, New York: John Wiley, 1985.

Kmenta, Jan, *Elements of Econometrics*, 2nd edition, New York: Macmillan, 1986.

Kravis, Irving B., Alan W. Heston, and Robert Summers, 'Real GDP *Per Capita* for More Than One Hundred Countries,' *Economic Journal*, June 1978, **88**: 215–242.

Kravis, Irving B., Alan Heston, and Robert Summers, *World Product and Income: International Comparisons of Real Gross Product*. Baltimore: Johns Hopkins University Press, 1982.

Kravis, Irving B., Alan W. Heston, and Robert Summers, 'The Share of Services in Economic Growth,' in F. Gerard Adams and Bert G. Hickman, eds, *Global Econometrics*. Cambridge, MA: The MIT Press, 1983, 188–218.

Kravis, Irving B., and Robert E. Lipsey, *Toward an Explanation of National Price Levels*, Princeton Studies in International Finance No. 52, Princeton: International Finance Section, Princeton University, 1983.

Kravis, Irving B., and Robert E. Lipsey, 'The Assessment of National Price Levels,' in

Sven W. Arndt and J. David Richardson, eds, *Real-Financial Linkages among Open Economies*, Cambridge, MA: The MIT Press, 1987, 97–134.

Kravis, Irving B., and Robert E. Lipsey, 'National Price Levels and the Prices of Tradables and Nontradables.' *American Economic Review*. May 1988, **78**: 474–478.

Officer, Lawrence H., *Purchasing Power Parity and Exchange Rates: Theory, Evidence and Relevance*, Greenwich. CT: JAI Press, 1982.

Officer, Lawrence H., 'The National Price Level: Theory and Estimation,' *Journal of Macroeconomics*. Summer 1989, **11**: 351–373.

Panagariya, Arvind, 'A Theoretical Explanation of Some Stylized Facts of Economic Growth.' *Quarterly Journal of Economics*, August 1988, **103**: 509–526.

Pincus, John, *Economic Aid and International Cost Sharing*. Baltimore: Johns Hopkins, 1965.

Summers, Robert, and Alan Heston, 'Improved International Comparisons of Real Product and Its Composition, 1950–1980,' *Review of Income and Wealth*, June 1984, **30**: 207–261.

Summers, Robert, and Alan Heston, 'A New Set of International Comparisons of Real Product and Price Levels Estimates for 130 Countries, 1950–1985,' *Review of Income and Wealth*, March 1988, **34**: 1–25.

Summers, Robert, Irving B. Kravis, and Alan Heston, 'International Comparisons of Real Product and Its Composition, 1950–77,' *Review of Income and Wealth*, March 1980, **26**: 19–66.

7 An assessment of the United Nations scale of assessments from a developing-country standpoint*

'Never have so many argued so much about so little money as in the United Nations.'

Stoessinger *et al.* (1964, p. 3)

Political scientists who are 'realists' (*e.g.* Claude, 1963; Stoessinger *et al.*, 1964; James, 1989) have long stated that financing the United Nations (UN) is a 'political' problem, in the sense that the geopolitical conflicts among countries outside the UN find intense focus within that body and in the context of its financing. Those observers who are 'idealists' (*e.g.* Padelford, 1963; Ogata and Volcker, 1993) take the view that if only countries would take the proper, international perspective, they would readily finance the UN and all would be well. Economists (*e.g.* Schelling, 1955; Kravis and Davenport, 1963; Pincus, 1965) have concentrated on the intricacies of proportional versus progressive taxation in international cost-sharing but paid minimal attention to the UN itself and, in any event, long ago opted out of the issue of UN financing.

This chapter examines the apportionment of UN expenses among its member states from a modern economic perspective. Unlike the political science literature, the focus is the regular budget rather than the peacekeeping accounts, because the distribution of peacekeeping expenses among members is derivative from the allocation of the regular budget. Regarding country-group friction in UN financing, political science studies consider East–West conflicts while the economic literature is oriented to tensions among the developed market economies themselves. In contrast, the focus of this chapter is the allocation of a expenses between developed and developing countries, which became the dominant issue in UN financing even before the Soviet Union began to erode internally and indeed since the late 1960s, if the nature of debate in the General Assembly is the criterion.

In Section I the country allocation of UN expenses is examined in a public choice context. The nature of UN 'output', the voting rules of the organization, and the peculiar type of allocation are shown to inhibit realization of the efficient amount of UN output as well as to exacerbate country-group conflict.

In Section II the outcome of such conflict is presented in the form of the allocation of UN expenses by country groups arranged according to level of development. While the main issue concerns the sharing of the financing burden between developed and developing countries, Eastern Europe and various groups of developing countries are also distinguished.

Section III examines just the initial phase in the UN procedure for country allocation of its expenses, because much of the allocation follows from that first step. The UN begins with the country distribution of nominal (exchange rate converted) net national product. Recent UN openness regarding its own income data, together with availability of new and improved real (purchasing-power-parity converted) income data in the public domain, permit an empirical evaluation of this first step. As economic theory predicts, the first step in the UN allocation procedure involves substantial reduction in the assessment of developing countries, because of the exchange-rate conversion bias.

In Section IV the remainder of the allocation process is shown again, on balance, to favor the developing countries, primarily because of the element of taxation progressivity. While this part of the allocation procedure is complex and multi-faceted, reflecting compromises among country groups, the outcome can be approximated remarkably well by converting the UN formula for progressivity to real income, dropping all other steps in the procedure, and explicitly incorporating level of development – done in Section V.

Section VI asks some questions raised by the study and offers some tentative answers. A data appendix follows.

I. Financing of UN expenses in a public-goods context

The United Nations Charter (Article I) states that the purposes of the United Nations are 'to maintain international peace and security . . . to develop friendly relations among nations . . . and to take other appropriate measures to strengthen universal peace . . . to achieve international co-operation in solving international problems of an economic, social, cultural, or humanitarian character . . . [and] to be a centre for harmonizing the actions of nations in the attainment of these common ends.'

Therefore the UN produces 'international' public goods for countries, just as a given country provides 'national' public goods to its citizens. Because of the free-rider problem, the UN cannot rely on voluntary contributions to finance its output. Therefore member states are taxed. This takes the form of a percentage allocation of UN expenses to the member states (the UN scale of assessments). It is a fundamental principle of the UN, often documented in General Assembly. (GA) resolutions, that ability to pay is the primary criterion for this allocation. A country's contribution to the UN regular budget is obtained as one hundredth the product of its assessment rate (or assessment) and the budget.[1] With assessment rates computed to two decimal places, a total of 100.00 percentage points, or 10,000 'points', are to be distributed over all countries.

Both the scale of assessments and the UN budget are decided in the GA by majority vote. To determine the scale, the GA acts on the recommendations of a Committee on Contributions (CC), but provides constraining directives to that body. A given scale is delineated by the scale period (the period for which the scale is applicable, currently three years) and the statistical base period (the period over which income data are averaged in the initial step of scale determination, currently ten years). Table 7.1 presents the attributes of the 1986–88, 1989–91 and 1992–94 scales, those scales for which data are available for the analysis of this paper.[2]

Given this method of allocating the burden of financing and given the GA base of UN output determination, free-riding can take some unusual forms. Firstly, in principle a country may elect not to join the UN, but in practice only some of the smallest countries may seek to achieve such a free ride. Secondly, realizing that determination of the scale of assessments is a zero-sum game, a country group may shift the tax burden onto another group simply by reducing its own assessment at the expense of the other group. This is the game that developing countries, the vast majority of UN membership, have played against developed countries since the 1960s.

Thirdly, a country group may insist on a quality or quantity of UN output that satisfies its own preferences but is antagonistic to those of another group, as the developing countries do in voting for development-oriented activities. Fourthly, a country may refuse to pay all or part of its assessment or may delay payment. Because the UN is not a true world government, enforcement of payment is ineffective.

Developed countries often complain that the UN budget is too high because of *X*-inefficiency and because of output that satisfies the preferences only of a select group (*i.e.* developing countries). Indeed, the foundation of the UN scale of assessments on ability to pay rather than the benefit principle

Table 7.1 Attributes of UN scales, 1986–1994

Scale period	*Statistical base period*	*UN membership*[a]	*Sum of assessments*
1986–88	1974–83	157	100.00
1989–91	1977–86	157	100.00
1992–94	1980–89	161[b]	100.02[c]

a Number of countries assessed at time scale was decided. Components of USSR (1986–91 scale periods), ex-USSR (1992–94 scale period), or ex-Yugoslavia (1992–94 scale period) counted as one country.

b Differs from membership for preceding scales by the inclusion of new members (Democratic People's Republic of Korea, Liechtenstein, Marshall Islands, Micronesia, Namibia, Republic of Korea) and the mergers of former members (German Democratic Republic and Federal Republic of Germany into Germany, People's Democratic Republic of Yemen and Yemen Arab Republic into Yemen).

c Differs from 100.00 because assessment rates for two new members (Marshall Islands and Micronesia, with assessment 0.01 each) were recommended in a CC addendum *Report*.

Source: CC Report (including addenda) and *FC Report*, various issues 40th sess. to 46th sess., 1985–1991.

divorces financing from expenditures and makes the achievement of an efficient level of output very difficult. Also, *X*-inefficiency is fostered by a country quota system for positions and a legislature (the GA) with ultimate authority but not itself elected. However, there are good reasons to believe that the scale of UN output is too *low* rather than too high.

Firstly, countries can react to a shift of the financing burden in their direction by going into arrears in their payment, and the knowledge of this possibility acts as a restraint on budget growth. Secondly, quite different from domestic tax systems, which are based on absolute income or expenditure, the UN scale of assessments involves no automatic increase in revenue as world inflation or indeed UN membership increases. Thirdly, the UN acts to produce peculiar kinds of public good – world peace and the solution of economic, social, cultural and humanitarian problems – that can substitute for the activities of national governments. Too effective a United Nations might involve a transfer of allegiance from individual countries to the UN itself and ultimately world government. Therefore it is in the interest of member governments to restrain the size and scope of UN activities.[3]

So it perhaps is not surprising that, by any reasonable standard, UN expenditure is low relative to its objectives. The total assessment of member states for both the regular budget and peacekeeping was $2.72 billion in 1992, less than the military spending of Thailand and less than the expenditure of New York City on its police and fire departments (Ogata and Volcker, 1993, pp. 3, 28, 32).

II. Grouping of countries by level of development

Country-group conflict is endemic in the GA and stimulated in UN financing by the zero-sum nature of the scale of assessments. Such conflict, especially between the developed and developing groups, exacerbates the problems of financing UN output and achieving an efficient level of that output. Therefore the scales in Table 7.1 are examined from the standpoint of six country groups: developed countries (DCs), Eastern Europe (EE), and four developing-country groups: Asian and Latin American newly industrialized countries (NICs), Organization of Petroleum Exporting Countries (OPEC), least developed countries (LDCs), and other developing countries (ODCs). The division among the developed countries, Eastern Europe, and the consolidated developing-country groups follows the classification of the United Nations Statistical Office in its various publications.[4]

NICs are identified by extending a table found in O'Neill (1984, p. 711). Thirteen studies that provide independent lists of Asian and Latin American newly industrialized countries are identified.[5] Brazil, Singapore and South Korea are on all thirteen lists; Mexico is on twelve; and Argentina eight. No other developing country on these continents receives as many as six 'votes' (except for the non-members of the UN, Hong Kong and Taiwan). So the five aforementioned countries constitute the NICs.

The OPEC membership of thirteen countries, as is the list of NICs, is the same for all three scales. LDCs are the 'least developed countries' officially designated as such by the UN, so the grouping varies for the three scales. Therefore the set of countries constituting the ODCs also varies.

Table 7.2 presents the UN scale of assessments for the six mutually exclusive country groups.[6] About three-quarters of UN expenses are borne by DCs and at least half the remainder by EE, leaving only a little over 10 per cent as the developing-country share (NICs, OPEC, LDCs and ODCs combined). Yet developing countries constitute over three-quarters the membership of the UN. Under a poll tax, the developing rather than developed countries would pay three-quarters of UN expenses. Under a crude application of the benefit principle, smaller and developing countries – with less ability to defend themselves – benefit more than large and developed countries from the UN role in fostering world peace and international cooperation. This criterion would suggest that the developing-country groups pay more than three-quarters of UN expenses!

From this simple analysis, it does appear that the developing countries may have succeeded in minimizing their joint assessment at the expense of DCs

Table 7.2 Various scales of assessment, by country group, 1986–1994 (percent)[a]

Period	*DCs*	*EE*	*NICs*	*OPEC*	*LDCs*	*ODCs*
			UN scale[b]			
1986–88	74.33	15.07	3.01	3.62	0.29	3.68
1989–91	74.45	14.63	3.16	3.77	0.39	3.60
1992–94	76.20	12.41	3.85	3.67	0.40	3.49
			Nominal NNP scale			
1986–88	67.81	11.81	4.68	5.33	0.63	9.73
1989–91	69.19	10.56	4.42	5.66	0.68	9.49
1992–94	71.59	8.78	4.99	4.75	0.66	9.23
			Real GNP scale			
1986–88	54.96	11.84	6.35	5.93	1.36	19.56
1989–91	54.38	11.50	6.23	5.70	1.68	20.50
1992–94	53.49	10.26	6.81	5.58	1.51	22.36
			Alternative scale			
1986–88	73.13	15.58	3.22	3.31	0.21	4.56
1989–91	72.42	15.32	3.52	3.28	0.26	5.20
1992–94	72.70	13.95	3.99	3.06	0.24	6.06

a Due to rounding, a row may not sum to 100.00.
b Row for 1992–94 sums to 100.02 by UN decision. See Table 7.1.

Source: UN Scale – 1986–88: *CC Report*, 40th sess., 1985, pp. 19–22; 1989–91: *CC Report*, 44th sess., 1989, pp. 20–27; 1992–94: *CC Report*, 46th sess., 1991, pp. 29–36. *Nominal NNP Scale* – 1986–88: *CC Report*, 45th sess., 1990, pp. 45–50 (published scale, which sums to 99.95, is multiplied by 100/99.95); 1989–91: *CC Report*, 44th sess., 1989, pp. 20–27; 1992–94: *CC Report*, 46th sess., 1991, pp. 29–36. *Real GNP Scale* and *Alternative Scale* – see Appendix and Sections IV, V.

and perhaps EE. Also, it is intriguing that the assessments of the 'richer' developing-country groups (NICs, OPEC, ODCs) are roughly of equal magnitude, while the assessment of LDCs, the poorest country-group in the world, is quite low. The richer developing-country groups, in effect, may have combined with the LDCs to minimize the over-all developing-country assessment and then shared most of that assessment almost equally, while accommodating the sorry economic conditions in LDCs.

However, it is ability to pay that is the primary principle of the UN scale of assessments. Have the developing-countries succeeded in using their overwhelming voting strength in the GA to shift to DCs (and perhaps EE) a disproportionate share of UN financing even under the criterion of ability to pay? Sections III and IV provide answers to this question.

III. The first step in determining the UN scale: nominal NNP versus real GNP

The CC obtains the first step in the generation of the UN scale as follows. For each member state, annual net national product (NNP) in current prices is multiplied by a conversion factor – generally the dollar/domestic currency exchange rate – and averaged over the statistical base period.[7] The resulting percentage distribution may be called the nominal NNP scale and is shown by country group in Table 7.2 for the three scale periods of the study.[8]

The nominal NNP scale would constitute a proportional tax on member states under the ability-to-pay criterion providing that converting to a common currency has a neutral effect. However, this would be so only were the conversion factor absolute purchasing power parity (PPP), which has been consistently and explicitly rejected by the CC. The CC evinces no awareness of the exchange rate conversion bias: the positive correlation of the PPP/exchange-rate ratio (both numerator and denominator expressed as number of units of domestic currency per dollar) with per capita income.[9] The exchange rate conversion bias implies that income in a common currency is biased downward for lower-income (meaning lower per capita real income) relative to higher-income countries when nominal (exchange rate converted) rather than real (PPP converted) income is considered; therefore it makes the scale progressive with respect to per capita *real* income, albeit proportional with respect to nominal income. Thus the nominal NNP scale, the starting point for determining the UN scale, immediately biases the outcome in favor of lower assessments for the developing-country groups.

The CC also errs in two other respects. Gross national product (GNP) rather than NNP would enhance international comparability of income, because accounting depreciation varies with tax law and data quality. Also, GNP is better measured in constant rather than current prices, because pure price change does not alter ability to pay.

So the NNP scale may be evaluated by comparing it with an alternative starting point for UN scale determination: the real GNP scale in constant prices. The latter scale emanates from an annual real GNP series in constant prices constructed for each UN member for which the requisite data exist, and then taking the average for the statistical base period. Although data are unavailable for the entire UN membership, the real GNP scale is computable by country group for the three periods by adjusting for differential sample coverage and is shown in Table 7.2. The scale is 'real' because the conversion factor for GNP in current domestic prices is PPP. It is 'in constant prices' because (i) the prices for PPP conversion pertain to a base year and (ii) income data for other years emanate from constant price national accounts.[10]

The real GNP scale is an appropriate scale for proportionate taxation. Rather than using this scale as the first step in UN scale determination, the CC employs the nominal NNP scale, which, as Table 7.2 shows, redistributes points from the developing-country groups and EE to DCs. The number of percentage points foisted on DCs by the use of nominal NNP rather than real GNP as the basis of the scale increases over time from almost 13 in 1986–88 to over 18 in 1992–94. The prime beneficiary is the ODCs, whose loss of percentage points increases from below 10 to over 13.

The conceptual divergences of nominal NNP from real GNP here are exclusion of depreciation, denomination in current rather than constant prices, and exchange rate rather than PPP conversion. The effect of the depreciation element on country-group assessments is uncertain. Expressing income in current rather than constant prices increases (reduces) the measured income of countries experiencing terms-of-trade improvement (deterioration) and therefore acts to increase (reduce) the country's assessment. On balance, this factor is probably favorable for the developing-country groups relative to DCs, the principal exception being OPEC under oil-price increases. Exchange rate conversion carries with it the exchange rate conversion bias, which definitely is favorable for ODCs and LDCs, as they are low-income relative to the other country groups, and unfavorable for DCs.

IV. Remaining steps in UN scale determination

After nominal NNP, the remaining steps in UN scale determination are as follows. Repayment of foreign debt principal is subtracted from NNP – though only imperfectly, because of measurement problems. This adjustment is advantageous to developing countries.[11] A component involving even greater progressivity in taxation (and therefore favoritism to the developing-country groups except perhaps OPEC) is a progressive tax element, which reduces income (and therefore the assessment) of low-income countries according to a two-parameter formula. Let G be the 'gradient' (the maximum percentage reduction, 85 per cent for the scales in Table 1), L the per capita income limit ($2200 for 1986–91, $2600 for 1992–94) and YP_i per capita NNP of country i. Debt adjusted income is decreased if $YP_i < L$, the percentage reduction given by $G \cdot [(L - YP_i)/L]$. Because income is expressed in nominal rather than real terms, progressivity and the advantage to developing countries are reinforced.

There is a 'ceiling' (maximum assessment, 25 per cent and effective only for the USA) and a 'floor' (minimum assessment, 0.01 per cent and effective for about half UN membership). Though clearly regressive and counter to the principle of ability to pay, these elements are not particularly important quantitatively. There are also scale-to-scale constraints: the assessment rates of LDCs are not to increase (ineffective in the 1986–94 scales) and there is a complex scheme limiting assessment changes for any country. Finally, the CC itself uses its judgment to make subjective reductions in the assessments of particular countries. The steps are applied in a particular sequence (and in fact the final scale would differ depending on the order) and at each step a predetermined method is used to reallocate the points added to or subtracted from the affected countries' assessments.

The effect of this complex procedure in transforming the nominal NNP scale to the UN scale is to shift still more points away from the developing-country groups – beyond the reallocation of points from the use of nominal NNP rather than real GNP – primarily because of the progressive tax element (see Table 7.2). The principal beneficiary is the ODCs, who enjoy an assessment reduction in the order of 5¾ to 6 percentage points. The points shed by the developing-country groups are borne by the DCs and EE. The absolute burden of the DCs is larger, but the EE *relative* burden – measured by the ratio of the UN to the nominal NNP scale assessment – is greater. For the three scale periods, this ratio is in the range 1.06–1.08 for DCs, 1.28–1.41 and increasing over time for EE. ODCs obtain the greatest relative advantage among the developing-country groups (the ratio steady at 0.38 over the three periods), LDCs receive the second greatest advantage but decreasing over time (the ratio at 0.46–0.61), while NICs and OPEC have similar ratios (0.64–0.77).

V. A sensible alternative to the UN scale

Determination of the UN scale is flawed because it begins with the nominal NNP scale rather than the real GNP scale. Its rationale is difficult to comprehend because the further steps in its determination are multi-faceted and complex. Also, the UN scale is disingenuous because it conceals favoritism toward the developing-country groups. This section develops an alternative to the UN scale that lacks these deficiencies and yet comes close to the UN scale outcome. The alternative scale can be simply described. It begins with the real GNP scale, adjusts per capita GNP for the level of development, and incorporates the progressive-tax element of UN scale determination in real-income terms. All the remaining elements and complications used in the UN scale determination are dropped.

There exists a literature on the multivariate use of socioeconomic indicators to measure level of development.[12] The work most pertinent to the present study is Berlage and Terweduwe (1988), who use factor analysis in effect to rate countries by level of development. Using their results, regression analysis transforms per capita real GNP to a level-of-development measure for all the developing countries for which the real GNP scale is available. Multiplying per capita real GNP by this measure adjusts income for development level. The UN progressive tax formula is then applied, with the UN parameter values for a given statistical base period, except that the per capita income limit is re-expressed in real income terms.

The resulting 'alternative scale' is shown in Table 7.2. Comparison with the real GNP scale reveals a substantial redistribution of points from the developing-country groups, especially ODCs, particularly to EE on a percentage basis and to DCs in absolute terms – even more so than when moving from the nominal NNP to the UN scale.

The alternative scale should be viewed as experimental, because the level-of-development measure is inherently arbitrary. Nevertheless, a comparison with the UN scale is instructive. The alternative scale is tantalizingly close to the UN scale, involving a shift of only about 1½ to 4¼ percentage points (albeit increasing over time) from DCs, OPEC and LDCs to the remaining groups. The UN takes a complex, flawed route to arrive at a scale that, on a country-group basis, is closely approximated by an alternative scale that is both simpler and methodologically pleasing, incorporating as it does both real income and explicit favoritism toward developing-country groups.

VI. Some questions and tentative answers

1. Why have economists neglected the topic of UN financing? They have chosen to leave the matter to the political scientists.

2. Have political scientists approached the issue satisfactorily? No, for two reasons. Firstly, they do not employ the concept of public goods. Secondly, they concentrate on the allocation of peacekeeping expenses rather than the regular budget.

3. Why is the UN scale so nakedly biased in favor of developing countries, by using exchange rates as the conversion factor for national income? It is not for lack of knowledge of the PPP alternative, as several times the CC has been informed of the International Comparison Project (which just happens to be lodged in the UN). The ostensible reason is data unavailability, which the Appendix suggests is not a defensible position. Probably, rejecting PPP is a hidden way of letting the developing-country groups off lightly in UN financing. With their overwhelming voting majority in the GA, the developing countries have arranged for their share of the burden of UN financing to fall below even their relative economic size.

4. Why is the alternative scale so close to the UN scale? The UN's complex and flawed procedure hides favoritism toward developing-country groups, which the alternative scale makes rational and explicit.

Appendix: Construction of scales

A. Real GNP scale

Step 1: From Summers and Heston (1991), take the annual series of per capita real GDP (international dollars, 1985 prices, Laspeyres index) and population (thousands of persons) for each UN member. What is desired are annual series for 1974–89 (1980–89 for the new members listed in note b of Table 7.1); but the Summers–Heston data stop in 1988, are available for an even more limited time period (sometimes only 1985) for some countries, and exclude certain members.[13] Nevertheless, except for their divergence from GNP, the income data are what is desired: the year 1985 emanates from a PPP computation and figures for other years are extrapolated using disaggregative national accounts series at constant prices.

Step 2: Compute real GDP in millions of international dollars as the product of population and per capita real GDP divided by 1000.

Step 3: Multiply real GDP by the GNP/GDP ratio (both numerator and denominator in current prices).[14] The result is real GNP in millions of international dollars.

Step 4: For cross-country consistency, the Summers–Heston data set fixes the country coverage. However, real GNP is extended backward (to 1974 or 1980, where needed) and forward to 1989 by means of series of domestic output at constant prices linked to real GNP by means of the overlap ratio at the earliest available year of real GNP for backward extrapolation and at the latest available year (generally 1988) for forward extrapolation.[15] The population series is extended by the same technique.[16]

Step 5: Adjust the real GNP and population series for mergers of countries (as indicated in note b of Table 7.1). For Germany in 1980–89, the respective series for the former Federal Republic of Germany and German Democratic Republic are summed for the 1992–94 scale period. However, for lack of data, Yemen is not in the sample for 1992–94.

Step 6: For each statistical base period delineated in Table 7.1, take the unweighted average of annual real GNP for each country i corresponding to U_i (where U_i is the UN assessment for country i, that is, the country enters the UN scale). The resulting real GNP is denoted as YA_i for country i (with the given statistical base period understood here and in all symbolic representations in the remainder of the Appendix).

Step 7: Let s denote the sample of countries for which YA_i is available. For each country group j (j = DCs, EE, NICs, OPEC, LDCs, ODCs), sum the YA_i and corresponding U_i separately: $YA^j = \Sigma YA_i$ and $U_s^j = \Sigma U_i$, $i \in j, s$. For each j, let $U^j = \Sigma U_i$ over all $i \in j$.

Step 8: Multiply each YA^j by the inverse of the group's proportionate sample coverage, (U^j/U^j_s), thus correcting for differences in sample coverage among the groups, resulting in YB^j.

Step 9: Multiply each YB^j by ($100/\Sigma YB^j$), yielding a scale summing to 100.00.[17]

B. Alternative scale

Step 1: For each country *i*, compute per capita real GNP (denoted by YQ_i) as the ratio of YA_i to mid-period population, where the denominator is consistent with CC practice.

Step 2: Berlage and Terweduwe (1988) perform a factor analysis of 18 socioeconomic variables over 102 countries, both developing and developed. They extract three factors, called 'level of a highly developed modern society', 'basic needs satisfaction', and 'industrial export-led growth', each of which may be construed as correlated with level of development. Letting F_i denote the weighted average of factor scores for country *i* (from Berlage and Terweduwe, 1988, p. 1537), use ordinary least-squares to fit the following regression over developing countries only ($i \in$ NICs, OPEC, LDCs, ODCs), for each scale period:

$$F_i = a + b \cdot YQ_i + c \cdot \text{NIC}_i + d \cdot \text{OPEC}_i + e \cdot \text{LDC}_i + u_i$$

where NIC_i (OPEC_i, LDC_i) is one if country *i* is a newly industrialized country (member of OPEC, least developed country), zero otherwise, and u_i is an independently and identically distributed error term. The regressions transform per capita real income (YQ_i) into a development level variable (F_i), allowing for country-group effects, and are shown in Table 7.3.

Step 3: Use the regression for a given scale period to convert per capita real

Table 7.3 Development-level equations

Scale period	*Constant*	*YQ*	*NIC*	*OPEC*	*LDC*	$\bar{R}^2$	*N*
1986–88	–0.39	0.00011	0.22	–0.10	–0.37		
	(9.28)	(8.41)	(1.91)	(1.04)	(6.23)	0.75	76
1989–91	–0.39	0.00011	0.15	–0.10	–0.32		
	(8.78)	(8.75)	(1.27)	(1.10)	(5.53)	0.75	75
1992–94	–0.38	0.00012	0.16	–0.08	–0.38		
	(8.73)	(8.91)	(1.58)	(0.96)	(6.80)	0.80	70

Dependent variables: *F* = development level, –1.00 (least developed) to 0.91 (most developed)
YQ = per capita real GNP, millions of international dollars
NIC = 1 if newly industrialized country, 0 otherwise
OPEC = 1 if member of Organization of Petroleum Exporting Countries, 0 otherwise
LDC = 1 if least developed country, 0 otherwise
$\bar{R}^2 = R^2$ corrected for degrees of freedom
N = number of observations

Source: see Appendix.

income (YQ_i) for all developing countries for which YA_i and therefore YQ_i exist, not just the regression sample, into development level ($\hat{F}_i$, the fitted value of F_i).

Step 4: The transformation

$$T_i = [\hat{F}_i - \min(\hat{F}_i)]/[\max(\hat{F}_i) - \min(\hat{F}_i)]$$

results in a development level variable with range zero (least developed) to one (most developed) among the developing countries. Members of the DC and EE groups are assigned the value $T_i = 1$.

Step 5: Construct $YR_i = T_i . YQ_i$, thus adjusting per capita real income for level of development.

Step 6: Convert the per capita income limit of the UN progressive tax formula to its real-GNP equivalent. This is done by multiplying the UN limit for the base period by the ratio of US real GNP (YA_i for i = US) to US nominal NNP, where the denominator (like the numerator) is averaged over the statistical base period.[18]

Step 7: For each country calculate taxable income, YC_i, by applying the UN progressive tax formula with the converted income limit of step 6 to YR_i.

Step 8: Sum the YC_i into country groups: $YC^j = \Sigma YC_i$, $i \in j, s$.

Step 9: Adjust the YC^j for differential sample coverage by multiplying each by the respective (U^j/U^j_s), denoting the result as YD^j.

Step 10: Multiply the YD^j by $(100/\Sigma YD^j)$, resulting in a scale summing to 100.00.

Notes

* Reprinted from *Journal of International Money and Finance*, vol. 13, Lawrence H. Officer, 'An Assessment of the United Nations Scale of Assessments from a Developing-Country Standpoint,' pages 415–428, Copyright 1994, with permission from Elsevier.

1 For peacekeeping expenses, a special 'peacekeeping scale', derived from the scale of assessments, is used. Least developed countries pay only 10 per cent and the remaining developing countries 20 per cent of their regular budget assessment rate, with the permanent members of the Security Council taking up the shortfall. For institutional information about the scale of assessments, see United Nations, General Assembly, *Report of the Committee on Contributions* (cited as *CC Report*) and *Scale of Assessments for the Apportionment of the Expenses of the United Nations: Report of the Fifth Committee* (cited as *FC Report*), annual per General Assembly session, plus occasional addenda; United Nations, General Assembly, *Evolution of the Methodology for the Scale of Assessments and its Current Application*, 49th sess., 24 April 1989; United Nations, *Yearbook of the United Nations*, annual; and US Department of State, *United States Participation in the United Nations*, annual.

2 An economic history of all UN scales since the inception of the organization is provided in Officer (1992).

3 For a milder argument along these lines, see Singer (1961, p. 178).

4 *International Trade Statistics Yearbook, Monthly Bulletin of Statistics*, and

Statistical Yearbook. DCs consist of the members of the Organization for Economic Co-operation and Development (OECD) excluding Turkey, plus Israel, Liechtenstein, Malta, and South Africa. EE comprises Albania, Bulgaria, Czechoslovakia, German Democratic Republic (1986–91 scales), Hungary, Poland, Romania, and the UN members comprising the USSR (1986–91 scales) or devolving from the ex-USSR (1992–94 scale).

5 The studies are the nine listed in O'Neill (1984, p. 712) plus Balassa (1998), Jenkins (1991), Ranis and Orrock (1985), and Turner and McMullen (1982).

6 The UN scale is formally determined on an individual country basis. Aggregation by groups is done to make the scale meaningful for the country-group orientation of this study.

7 A subsidiary conversion factor is relative purchasing power parity, called PARE (price-adjusted rate of exchange) by the CC. A base year is selected, and the current year conversion factor is the product of the base year exchange rate and the domestic/US current-to-base-year price-index ratio. Though the CC considers PARE to be the ideal conversion factor, its use is rare, confined to seven countries for the 1992–94 scale and not at all employed in the 1986–91 scales (though World Bank dollar denominated income data, that employ a kind of relative PPP concept, are utilized for three countries in the 1986–88 scale).

8 These scale periods are delimited by the fact that the CC did not reveal its nominal NNP scale until the 1986–88 period.

9 The evidence for the positive correlation of the PPP/exchange rate ratio with per capita income is threefold. Firstly is the 'causal empirical observation that services and nontradables generally are relatively cheap in low-income countries' (Kravis and Lipsey, 1983, p. 11). Secondly, at least seven economic theories predict the directional correlation: (i) a productivity-differential model, (ii) a factor-proportions model, (iii) a specific-factors model, (iv) an economies-of-scale model, (v) a financial-development model, (vi) a demand-oriented model, and (vii) a general-equilibrium model in which both the PPP/exchange rate ratio and per capita income are endogenous variables; presented well in Kravis and Lipsey (1983, pp. 4–5, 11–12), Bhagwati (1984), Clague (1985), Panagariya (1988), Feldman (1991), Bergstrand (1991), and Falvey and Gemmell (1991), respectively. Thirdly, a set of econometric investigations supports the correlation: Clague (1986, 1988a, 1988b), Falvey and Gemmell (1991), Kravis *et al.* (1983), Kravis and Lipsey (1983, 1987, 1988), Officer (1989), and Summers and Heston (1991).

10 Details on construction of the real GNP scale and the 'alternative scale' (developed in Section V) are in the Appendix.

11 It may be noted that it is inappropriate accounting to include a capital account transaction in current income. Also, there is asymmetry in (i) subtracting repayment of principal from debtor's income while not adding it to creditor's income and (ii) accounting for repayment but not the original borrowing.

12 Cluster analysis, discriminant analysis, factor analysis, and principal-components analysis are techniques utilized in the various studies. See, for example, Ram (1982), Dellaportas (1983), Khan and Zerby (1984), and Berlage and Terweduwe (1988).

13 For countries in their table B3, per capita GDP for 1985 is obtained as the product of 0.01, relative per capita GDP, and US 1985 per capita GDP.

14 The data source for the ratio is World Bank, *World Tables*, various issues.

15 While the extrapolation method is aggregative, it is consistent with the Summers-Heston income concept. Data sources for the output series are Alton *et al.* (1980, 1990); Central Intelligence Agency, *Handbook of Economic Statistics*, various issues; Summers and Heston (1988); and World Bank, *World Tables*, various issues.

16 Data sources are UN, *Demographic Yearbook*, and World Bank, *World Tables*, various issues.

17 The percentage sample coverage for the real GNP and alternative scales is ΣU_s^j, and is extremely high: 99.66, 99.52 and 99.13 for 1986–88, 1989–91 and 1992–94, respectively.

18 The source of US nominal NNP is US Department of Commerce (1986) and *Survey of Current Business*, various issues.

References

Alton, Thad P. *et al., Economic Growth in Eastern Europe 1965–1979*, Occasional Paper No. 59, Research Project on National Income in East Central Europe, New York: L.W. International Financial Research, 1980.

Alton, Thad P. *et al., Economic Growth in Eastern Europe 1975–1989*, Occasional Paper No. 110, Research Project on National Income in East Central Europe, New York: L.W. International Financial Research, 1990.

Balassa, Bela, 'The Lessons of East Asian Development: An Overview,' *Economic Development and Cultural Change*, April 1988, Supplement, **36**: S273–S290.

Bergstrand, Jeffrey H., 'Structural Determinants of Real Exchange Rates and National Price Levels: Some Empirical Evidence,' *American Economic Review*, March 1991, **81**: 325–334.

Berlage, Lodewijk, and Dirk Terweduwe, 'The Classification of Countries by Cluster and by Factor Analysis,' *World Development*, December 1988, **16**: 1527–1545.

Bhagwati, Jagdish N., 'Why are Services Cheaper in the Poor Countries?' *Economic Journal*, June 1984, **94**: 279–286.

Clague, Christopher K., 'A Model of Real National Price Levels,' *Southern Economic Journal*, April 1985, **51**: 998–1017.

Clague, Christopher K., 'Determinants of the National Price Level: Some Empirical Results,' *Review of Economics and Statistics*, May 1986, **68**: 320–323.

Clague, Christopher K., 'Explanations of National Price Levels,' in Jorge Salazar-Carrillo and D.S. Prasada Rao, eds, *World Comparison of Incomes, Prices and Product*, Amsterdam: North-Holland, 1988, 237–262 (1988a).

Clague, Christopher K., 'Purchasing-Power Parities and Exchange Rates in Latin America,' *Economic Development and Cultural Change*, April 1988, **36**: 529–541 (1988b).

Claude, Inis L., Jr., 'The Political Framework of the United Nations' Financial Problems,' *International Organization*, Autumn 1963, **17**: 831–859.

Dellaportas, George, 'Classification of Nations as Developed and Less Developed: An Arrangement by Discriminant Analysis of Socioeconomic Data,' *American Journal of Economics and Sociology*, April 1983, **42**: 153–166.

Falvey, Rodney E., and Norman Gemmell, 'Explaining Service-Price Differences in International Comparisons,' *American Economic Review*, December 1991, **81**: 1295–1309.

Feldman, David H., 'Economic Policy and the Relative Price of Services in LDCs,' *World Development*, October 1991, **19**: 1381–1389.

James, Alan, 'The Security Council: Paying for Peacekeeping,' in David P. Forsythe, ed, *The United Nations in the World Political Economy*, New York: St. Martin's Press, 1989, 13–35.

Jenkins, Rhys, 'The Political Economy of Industrialization: A Comparison of Latin American and East Asian Newly Industrializing Countries,' *Development and Change*, April 1991, **22**: 197–231.

Khan, M. H., and J. A. Zerby, 'An International Comparison of Socioeconomic Development,' *Scandinavian Journal of Development Alternatives*, March 1984, **3**: 25–37.

Kravis, Irving B., and Michael W. S. Davenport, 'The Political Arithmetic of International Burden-Sharing,' *Journal of Political Economy*, August 1963, **71**: 309–330.

Kravis, Irving B., Alan W. Heston, and Robert Summers, 'The Share of Services in Economic Growth,' in F. Gerard Adams and Bert G. Hickman, eds, *Global Econometrics*, Cambridge, MA: The MIT Press, 1983, 188–218.

Kravis, Irving B., and Robert E. Lipsey, *Toward an Explanation of National Price Levels*, Princeton Studies in International Finance No. 52, Princeton: International Finance Section, Princeton University, 1983.

Kravis, Irving B., and Robert E. Lipsey, 'The Assessment of National Price Levels,' in Sven W. Arndt and J. David Richardson, eds, *Real-Financial Linkages Among Open Economies*, Cambridge, MA: The MIT Press, 1987, 97–134.

Kravis, Irving B., and Robert E. Lipsey, 'National Price Levels and the Prices of Tradables and Nontradables,' *American Economic Review*, May 1988, **78**: 474–478.

Officer, Lawrence H., 'The National Price Level: Theory and Estimation,' *Journal of Macroeconomics*, Summer 1989, **11**, 351–373.

Officer, Lawrence H., *An Assessment of the United Nations Scale of Assessments, 1946–1994*, paper prepared for Conference on International Comparison of Price Structures, and their Influence on Productivity and Growth, July 17–18, 1992.

Ogata, Shijuro, and Paul Volcker, co-chairmen, *Financing an Effective United Nations: A Report of the Independent Advisory Group on U.N. Financing*, Ford Foundation, 1993.

O'Neill, Helen, 'HICs, MICs, NICs and LICs: Some Elements in the Political Economy of Graduation and Differentiation,' *World Development*, July 1984, **12**: 693–712.

Padelford, Norman J., 'Financial Crisis and the Future of the United Nations,' *World Politics*, July 1963, **15**: 531–568.

Panagariya, Arvind, 'A Theoretical Explanation of Some Stylized Facts of Economic Growth,' *Quarterly Journal of Economics*, August 1988, **103**: 509–526.

Pincus, John, *Economic Aid and International Cost Sharing*, Baltimore: Johns Hopkins Press, 1965.

Ram, Rati, 'Composite Indices of Physical Quality of Life, Basic Needs Fulfilment, and Income: A "Principal Component" Representation,' *Journal of Development Economics*, October 1982, **11**: 227–247.

Ranis, Gustav, and Louise Orrock, 'Latin American and East Asian NICs: Development Strategies Compared,' in Esperanza Durán, ed, *Latin America and the World Recession*, Cambridge: Cambridge University Press, 1985, 48–66.

Schelling, Thomas C., *International Cost-Sharing Arrangements*, Essays in International Finance No. 24, Princeton: International Finance Section, Princeton University, 1955.

Singer, J. David, *Financing International Organization: The United Nations Budget Process*, The Hague: Martinus Nijhoff, 1961.

Stoessinger, John G., *et al., Financing the United Nations System*, Washington, DC: The Brookings Institution, 1964.

Summers, Robert, and Alan Heston, 'A New Set of International Comparisons of Real Product and Price Levels Estimates for 130 Countries, 1950–1985,' *Review of Income and Wealth*, March 1988, **34**: 1–25.

Summers, Robert, and Alan Heston, 'The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950–1988,' *Quarterly Journal of Economics*, May 1991, **106**: 327–368.

Turner, Louis, and Neil McMullen, *The Newly Industrializing Countries: Trade and Adjustment*, London: Royal Institute of International Affairs, George Allen & Unwin, 1982.

US Department of Commerce, *The National Income and Product Accounts of the United States, 1929–82: Statistical Tables*, Washington, DC: US Government Printing Office, 1986.

8 An assessment of the United Nations scale of assessments, 1946–1994*

Never have so many argued so much about so little money as in the United Nations.

Stoessinger (1964, p. 3)

1. Concept and application of scale of assessments

The focus of this study is the UN scale of assessments, the percentage allocation of UN expenses to its members.[1] Each member state is assigned an assessment rate or assessment, computed to two decimal places, and the sum of the assessments over all countries is typically 100.00 percentage points or, in UN terminology, 10,000 points (a point being one-hundredth of a percentage point).

Two implications follow. First, determination of the scale is a zero-sum game, a clear invitation to country conflict. Second, the amount (generally required to be in dollars) contributed by a country is the product of three factors: (1) the UN budget, (2) one-hundredth the country's assessment rate, and (3) the proportion of the amount due to the United Nations [the product of (1) and (2)] that the country elects to contribute. A country dissatisfied with its assessment rate (or with the UN budget or the activities financed by the budget) can elect a low third factor (below unity) and become in arrears in its UN contribution.[2]

2. Selection of scales for study

The objective of this study is to present a history of the UN scale of assessments from three standpoints: first, accepting both the scale and the procedure used by the UN in setting the scale, and analyzing the steps in this procedure (Sections 3, 4 and 5); second, accepting the scale itself but simplifying and reinterpreting the procedure via an econometric model (Sections 6 and 7); third, rejecting both the UN procedure and the resultant scale, developing an 'ideal' scale in its place (Sections 8 and 9). The techniques of the second and third standpoints involve moving cross-section analysis over time. Therefore a time series of scales must be identified and selected.

According to the UN Charter, the General Assembly (GA) is responsible for apportioning the expenses of the UN among its members. The GA appoints a Committee on Contributions (CC) to make recommendations concerning that task. Primary elements of the CC's recommendations are the assessment scale itself and the scale period (the year or years for which the scale is applicable), with a three-year scale period envisaged in the GA's rules of procedure. Underlying the CC's recommendations is the 'statistical base period', the period for the statistical data used by the CC to determine the scale. In spite of a typical condition of contentious GA debate on any recommended scale, only rarely does that body alter the recommendations of the CC, due largely to the fact that guidelines for the CC's work are set by the GA itself.

Consider the concept of the decision period, the year in which the CC recommends a scale and the GA decides to establish a scale. A decision period yields the pertinent scale of assessments, the scale period, and the underlying statistical base period. The methodology for selecting the time series of scales for this study is as follows:

1. *Consider only decision periods at which all member countries are considered jointly*. The principal reason is that scales are then determined in a pure general-equilibrium setting, with uniform treatment of countries. New members, added between selected decision periods, are excluded until the next decision period. A subsidiary reason for this rule is that the sum of assessments generally totals the original 100.00 percentage points, instead of a higher number derived from adding the assessments of new members to the original number.
2. *Where more than one scale utilizes a given statistical base period, select the scale with the larger number of countries (always the later scale).* This maximizes the size of the country sample.
3. *Exclude scale periods where changes in the scale are trivial* (1948 and 1949).

The resulting time series of quintuples – decision period, scale period, statistical base period, and (representing the scale) UN membership and sum of assessments – are listed in Table 8.1. The scales themselves are found in the source to the table.

3. Properties of components of the scale

The elements of the scale are the steps in the procedure of determining the scale. They may be characterized in three different ways. First, a scale element may be objective or subjective in nature. The objective elements are based on statistical data, involve a rigid procedure and order of application, and are completely performed prior to application of the subjective elements. The

Table 8.1 Delineation of scale periods, 1946–1994

Decision period (year)	*Scale period*	*Stat. base period*[a]	*Norm. base period*[b]	*UN membership*[c] *Total*[d]	*UN membership*[c] *Floor*[e]	*Sum of assessments*
1946	1947[f]	1938–40	1938–40	52	7	100.00
1949	1950[g]	1948[h]	1948	57	8	100.00
1950	1951	1949[h]	1949	57	8	100.00
1951	1952	1950[h]	1950	58	8	100.00
1952	1953	1950–51	1950–51	58	9	100.00
1953	1954	1950–52	1950–52	58	9	100.00
1954	1955	1951–53	1951–53	58	9	100.00
1957	1958[i]	1952–54	1952–54	80	14	100.00
1958	1959–61	1955–57	1955–57	79[j]	16	100.00
1961	1962–64	1957–59	1957–59	98[k]	32	100.00[l]
1965	1966–67[m]	1960–62	1960–62	115[n]	51	99.82
1967	1968–70	1963–65	1963–65	120	57	100.00
1970	1971–73	1966–68	1966–68	124	63	100.00
1973	1974–76	1969–71	1969–71	133	70	100.00
1976	1977	1972–74	1972–74	142	81	100.00
1977	1978–79	1969–75	1973–75	145	66	100.00
1979	1980–82	1971–77	1975–77	149	70	100.00
1982	1983–85	1971–80	1978–80	155	78	100.00
1985	1986–88	1974–83	1981–83	157	78	100.00
1988	1989–91	1977–86	1984–86	157	79	100.00
1991	1992–94	1980–89	1987–89	161	84	100.02

Source: CC Report and FC Report, 1st – 46th sessions, 1946–1991.

a Statistical base period, used by UN and in sections 2–7 of text.
b Normative base period, used in sections 8–9 of text.
c Number of countries assessed at time scale was decided. Excludes new members assessed after decision period.
d Components of USSR (scale periods 1946–1991) or ex-USSR (scale period 1992–94) counted as one country.
e Number of countries assessed at floor rate.
f Scale for 1946 decided at same time with same statistical base period but three fewer countries.
g Scale for 1948 differs from 1947 by inclusion of two new members with the sum of their assessments (0.31) balanced by reduction in the assessment for Sweden. Scale for 1949 differs from 1948 by inclusion of one new member with its assessment of 0.15 balanced by reduction in assessment for Sweden of 0.04 and for United Kingdom of 0.11. Scale for 1950 differs from 1949 by division of assessment for 'India and Pakistan' (formerly assessment for India, unchanged since scale for 1947) into assessments for India and Pakistan separately, and by inclusion of one new member with its assessment of 0.12 balanced by reduction in assessment for Sweden of 0.02 and for United States of 0.10.
h Actually, the latest year for which income data are available for a country, of which probably the modal year.
i Scale for 1956–57 (decided in 1956) has same statistical base period but six fewer countries.
j United Arab Republic, joining Egypt and Syria, reduced UN membership by one.
k Syria and United Arab Republic (Egypt) counted as two members again. Necessity for allocation of assessment of former joint United Arab Republic to these two members noted in 1961 decision period, though not established until 1962.
l Excludes retroactive downward revisions in assessments for Czechoslovakia and Hungary, decided in 1963.
m Scale for 1965 decided at same time with same statistical base period but three fewer countries. Differs from 1966–67 scale by a higher assessment for Malaysia and exclusion of the three new members (one of which, Singapore, was then part of Malaysia).
n Excludes Indonesia, which temporarily withdrew from the United Nations in January 1965 (before the 1965 decision period).

subjective elements are based on the judgment of the CC and the order of their application is irrelevant (except that, beginning with the 1992–94 scale, the subjective elements became dependent entirely on the voluntary acceptance of additional points by donor countries).[3]

Second, a scale element corresponds to one or more assessment principles embedded in GA resolutions, even though the GA does not always make this correspondence explicit. Ability to pay is and always has been the fundamental principle for the scale. In a resolution of February 13, 1946, the GA stated: 'The expenses of the United Nations should be apportioned broadly according to capacity to pay.' A GA resolution of December 21, 1991 (among others over the years) reaffirms that 'the capacity of Member States to pay is the fundamental criterion for determining the scale of assessments.'

Three other principles of assessment have been noted by the GA. Collective financial responsibility is propounded in a GA resolution of December 14, 1976. Level of development is a principle present in many GA resolutions, perhaps best stated as the requirement of drawing up scales 'on the basis of . . . the continuing disparity between the economies of developed and developing countries' (GA resolution of December 14, 1976). Limited assessment change is a principle described as 'the need to prevent extreme and excessive variations of individual rates of assessments between two successive scales' in a GA resolution of December 18, 1981 (and similarly elsewhere).

Third, a scale element may have the effect of making the scale progressive, proportional, or regressive – but in two distinct senses. Let A_i denote the assessment of country i, where typically $\Sigma A_i = 100.00$, YS_i the income of country i as a percentage of total income over all countries (so that $\Sigma YS_i = \Sigma A_i$), and YP_i the per capita income of country i (leaving aside the issues of income definition and comparison across countries). Then according as the effect of the element is $\partial A_i / \partial YS_i$ greater than, equal to, or less than unity, the scale is progressive, proportional, or regressive with respect to total income and that element. Also, according as the effect of the element is $\partial A_i / \partial YP_i$ greater than, equal to, or less than zero, the scale is progressive, proportional, or regressive with respect to per capita income and that element. The two senses of progressivity are not distinguished in the existing literature.

4. History of the scale elements

Consider first the objective scale elements, arranged according to assessment principle, beginning with *ability to pay*. For the early scales the income concept was net national product (**NNP**) at factor cost; from the 1965 scale onward it was **NNP** at market prices. **NNP** as an income concept makes the scale proportional with respect to total income; for if the procedure were to stop right here, the scale would be identical to the percentage distribution of **NNP** (what the UN calls the national-income scale, but is better described as the **NNP** scale). From the 1986–88 scale onward, **NNP** was replaced by debt-adjusted income, obtained by subtracting repayment of foreign-debt principal (with interest payments already excluded from **NNP**), this element probably making the scale progressive with respect to per capita income.

In every scale, the income of poor countries has been reduced via a low-per-capita-income allowance, making the scale progressive with respect to per capita income. The percentage reduction of income is given by the formula $G[(L - YP_i)/L]$, where G is the gradient (the maximum percentage reduction) and L the per-capita-income limit (income is reduced only if $YP_i < L$). Over the years, G and L have been increased from the original 40 percent and $1000 to 85 percent and $2600 in the 1992–94 scale.

For each country, income in current (not constant) prices is averaged over the statistical base period; but, first, annual income must be expressed in a common currency, the U.S. dollar. The principal conversion factor is the current exchange rate (**ER**), in annual-average form. A subsidiary conversion factor is relative purchasing power parity: a base year is selected, and the current-year conversion factor is the product of the base-year exchange rate and the domestic/U.S. current-to-base-year price-index ratio. Eventually called **PARE** (price-adjusted rate of exchange) by the CC and viewed by that body as the ideal conversion factor, its use to date is rare: an unknown, but clearly small, number of cases in early scales and seven countries in the 1992–94 scale.[4]

The CC has considered absolute purchasing power parity (**PPP**) as a conversion factor, but rejected it, primarily for perceived lack of data. Surprisingly, it has evinced no awareness of the exchange-rate conversion bias (**ERCB**), well-known in the economic literature: the positive correlation of **PPP/ER** (both numerator and denominator expressed as number of units of domestic currency per dollar) with per capita income, that implies income is biased downward for lower-income relative to higher-income countries when estimated by nominal (**ER**-converted) income rather than expressed directly in real (**PPP**-converted) terms. The UN **NNP** scale is best termed the nominal **NNP** scale, the percentage distribution of the average of annual nominal **NNP** over the statistical base period. The **ERCB** transforms the proportionality of the nominal **NNP** scale with respect to total income into progressivity of the corresponding real **NNP** scale with respect to per capita income.[5]

Turning to *collective financial responsibility*, this principle justifies a ceiling (maximum assessment), per-capita ceiling (maximum per-capita assessment, that of the country with the highest assessment), and floor (minimum assessment). The ceiling, with values for all scales shown in the second column of Table 8.5, has been applicable only to the United States. The floor, originally set at 0.04 percent, changed to 0.02 with the 1974–76 scale and to 0.01 with the 1978–79 scale, has also been in existence for all scales. As the 'UN Membership' columns of Table 8.1 show, in the order of half the members have been assessed at the floor since the 1968–70 scale. The per-capita ceiling, in existence for the 1951–1976 scales, was enjoyed by only a few countries. All three elements are regressive – the ceiling and floor with respect to both total and per capita income, the per-capita ceiling with respect to per capita income.

Level of development provides a second justification of the low per-capita income allowance and also underlies 'no increase in the assessments of the least developed countries' (the latter in effect since the 1983–85 scale). *Limited assessment change* relates to many elements over the years: minimal change in the scale (1948–50 and 1966–67 scales; see Table 8.1 for details), maximum of ten-percent change in a given assessment (1951 scale), partial removal of discrepancies from income adjusted for low per-capita-income allowance or from ceiling (1952–55 scales), a longer statistical base period (1978–), and a 'scheme of limits' (1986–) involving eight assessment brackets each with a maximum percentage change and a maximum percentage-point change.

The hypothetical assessment scale based only on the objective elements is called the *machine scale* and results from the *scale methodology*, that is, application of the elements in a specific order with predetermined methods to reallocate the points shifted by each element.[6]

Considering now the subjective elements of the scale, the process of applying them is called *mitigation of the scale* or *ad hoc adjustment*. Under ability to pay, mitigation typically corrects for data problems, such as the prewar statistical base period (1946–47 scales) and varying data quality among countries (1951–61 scales). Mitigation has also allowed for casualty losses (due to wars or natural disasters) and such elements as **PARE**, economic events beyond the base period, external debt, and ability to secure foreign exchange—without incorporation in the scale methodology.

Under *level of development*, increases of assessments for developing countries have been mitigated or downward adjustments in assessments made. Similarly, for *limited assessment change*, changes (especially increases) in assessments were mitigated prior to establishment of the scheme of limits. The actual scale of assessments, called the *official scale*, is obtained by mitigating the machine scale, that is, redistributing points in that scale by applying ad hoc adjustment.

5. Importance of the scale elements

The quantitative importance of an individual element in the scale is the number of percentage points redistributed among member states by the step (in the UN scale procedure) corresponding to that element *relative to the hypothetical scale* determined at the *end of the preceding step*. While other definitions are possible, this one respects both the UN scale procedure (the scale methodology plus ad hoc adjustment) and the UN scale itself, which is the viewpoint of Sections 3 to 5.

Rows (1) to (6) of Table 8.2 (inclusive of notes) present all that is known, or can be computed, regarding the importance of individual elements in the scales from 1946 to 1992–94. Until the 1989–91 scale, information is scattered, because the CC was traditionally secretive regarding quantitative aspects of its work.[7] Taking the nominal-NNP scale as the basis, rows (1) to (6), in order, completely specify the remainder of the UN scale methodology

Table 8.2 Importance of elements in scale[a] (no. of percentage points redistributed[b])

	Scale period			
Element	*1983–85*	*1986–88*	*1989–91*	*1992–94*
(1) Debt adjustment			.87	.71
(2) Low-per-capita-income allowance[c]	9.50		8.35	8.27
(3) Floor[d]			.47	.51
(4) Ceiling[e]			5.68	5.51
(5) Scheme of Limits		1.79	3.54	3.82
(6) Mitigation[f]	1.52	.69	.77	.50
(7) Net effect of (1)–(6)[g]	9.88	11.67	12.90	11.73

Source: Rows (1)–(6); 1989–91, 1992–94–CC Report, 44th sess., 1989, p. 34; 46th sess., 1991, p. 44. Row (2), 1978–79–CC Report, 32nd sess., 1977, pp. 19, 22; 1980–82–CC Report, 34th sess., 1979, p. 13; 1983–85–CC Report, 37th sess., 1982, p. 7. Row (4), 1962–64–FC Report, 17th sess., 1962, p. 2; 1968–70–CC Report, 24th sess., 1969, p. 19; 1971–73–FC Report, 27th sess., 1972, p. 3. Row (5), 1986–88–CC Report, 45th sess., 1990, p. 6. Row (6), 1971–73–Evolution, p. 15; 1986–88–CC Report, 45th sess., 1990, p. 56. Row (7) – see Section 5 of text.

Rows (3) and (4), 1946 and 1947: The CC did not apply the principle of collective financial responsibility to its initial, 1946–48, recommended scale, resulting in a U.S. assessment of 49.89 percent. The GA rejected the CC scale and imposed a ceiling of 39.89 percent, a floor of 0.04 percent, and one-year scale periods. The importance of the ceiling is clearly ten percentage points; that of the floor is computed as $\Sigma \mid A_i - B_i \mid$, where A_i is country i's assessment for 1947, B_i its assessment recommended by the CC, and i runs over all countries for which $B_i < 0.04$.

Row (6), 1983–85: The CC incorrectly calculates the importance of mitigation in 1983–85 as 0.63 percentage points (CC Report, 45th sess., 1990, p. 50). The true figure, 1.52 percentage points, is computed as $(\Sigma \mid A_i - M_i \mid)/2$, where A_i is country i's official assessment and M_i its machine-scale assessment (source for the latter, CC Report, 37th sess., 1982). The explanation for the CC's error is that in constructing the 1983–85 scale, two mitigation processes occurred; first, the usual one, to obtain the CC recommended scale from the machine scale; second, an additional 0.63 percentage points redistributed because the GA would not accept the recommended scale. The CC computation, incorrectly, used the original recommended scale instead of the machine scale.

a A blank space indicates the figure is not available. A dash indicates the element is not applicable.
b Basis is nominal NNP at factor cost (1946–64), nominal NNP at market prices (1965–94).
c Other scale periods: 1978–79, 5.81; 1980–82, 8.85.
d Other scale periods: 1946 and 1947, 0.14.
e Other scale periods: 1946 and 1947, 10.00; 1962–64, 6.48; 1968–70, 7.91; 1971–73, 6.88.
f Other scale period: 1971–73, 0.20–0.25.
g And of 'No increase in the assessments of the least developed countries.'

plus ad hoc adjustment, except for the exclusion of 'no increase in the assessments of the least developed countries.'[8]

Because the information for a given element in the scale is so scattered, it is hard to discern behavior over time; but something can be said for each element. Debt adjustment entered the machine scale in 1986–88. It is highly probable that, in spite of a different schemata, that element redistributed less than one percentage point in that period, as it did in the two subsequent periods. The low-per-capita-income allowance is clearly of substantial importance since 1978–79. For earlier scales, it is reasonable to surmise that its impact was somewhat less, because of the lower gradient in the formula.

An upper limit to the number of points redistributed by the floor is the product of the floor assessment and the number of floor countries (countries assessed at the floor, some of which would be there even without application of the floor constraint). For the four available scale periods in Table 8.2, the ratio of the floor's actual point redistribution to the upper limit is in the .50–.61 range. If a ratio of .6 is used, the maximum effect of the floor occurred in 1971–73, with importance of $63 \times .04 \times .6 = 1.51$ percentage points.

Of all the elements, the ceiling's importance in the missing periods is the most uncertain. The lower ceiling over time (to 1974–76) increases the effect, but the reduction in U.S. relative nominal income acts to reduce it. The available figures suggest that, on balance, the trend in the ceiling's importance is downward. Information about the scheme of limits is complete for its existence; its importance has been increasing over time. However, there is reason to believe the GA will decide to terminate the scheme of limits.

Except for the 1983–85 aberration, the importance of mitigation has had an upward trend from the early scales (see also *Evolution*, p. 15). Cumulative criticism in the GA against mitigation, combined with making the acceptance of points purely voluntary, decreased the importance of mitigation in 1992–94.

Absent the elements that follow nominal **NNP**, the official scale would be the nominal-**NNP** scale. Information on the latter scale exists only for the scale periods exhibited in Table 8.2. For the 1986–94 scales, the nominal-**NNP** scale itself is known; for the 1983–85 scale, what is available is only dollar-denominated **NNP** – denoted as NNP_i for country i— and only for countries with assessment above 0.03 percent.[9]

The net joint effect of all scale elements that follow nominal **NNP**, these elements represented by rows (1) to (6), is $(\Sigma \mid A_i - NNPS_i \mid) /2$, where A_i is country i's official-scale assessment and $NNPS_i$ its **NNP**-scale assessment. This computation is made directly for the 1986–94 scales; but for 1983–85 one must resort to a three-step process. First, for the sample, s1, of countries with $A_i > 0.03$, the hypothetical $NNPS_i$ for a UN membership of s1 is $NNP_i \cdot (\Sigma^* A_i / \Sigma^* NNP_i)$, where Σ^* denotes restricted summation, in this case over the set s1 of i. Second, the net joint effect of the scale elements following nominal **NNP** is $Es1 \equiv (\Sigma^* \mid A_i - NNPS_i \mid)/2$. Third, for the full UN membership the effect is estimated as $Es1(\Sigma A_i / \Sigma^* A_i)$, where, as usual, Σ denotes unrestricted summation, that is, over the set of countries constituting the entire UN membership. The underlying assumption is that the proportion of points redistributed (from the nominal-**NNP** scale) outside the sample (that is, for the $i \notin s1$) is the same as within it (that is, for the $i \in s1$).[10]

The resulting net joint effect of the scale elements following nominal **NNP**, these elements represented by rows (1) to (6), is shown in row (7). The effect, number of points redistributed from the nominal-**NNP** scale, has range approximately 10–13 percentage points for the 1983–94 scales, meaning that nominal **NNP** accounts for 87–90 per cent of the scale.

6. Country conflicts and groupings

Throughout UN history, the work of the CC and debate in the GA has been replete with country-group conflicts regarding the scale of assessment. Four issues can be discerned. First, the United States has always argued, from the principle of sovereign equality of UN members (in Article 2 of the UN Charter), for the existence of the ceiling element in the scale methodology and for reductions in the ceiling level. Opposition to the U.S. position, from the Soviet bloc and from developing countries, centered on incompatibility of the ceiling with the ability-to-pay principle and the unique benefits the United States derived from UN membership (for example, UN headquarters in New York, payment of assessments in U.S. dollars).

Second, in the 1950s and 1960s countries of the Soviet bloc complained vigorously about increases in their assessments, claiming that their heavy war damage was being ignored and that a systematic effort to reduce assessments of Western countries at Eastern European and USSR expense was occurring. The United States and others countered that these assessment shifts reflected changes in relative capacities to pay and application of the ceiling, the latter legitimate via the principle of sovereign equality. The Soviet bloc also asserted that improper data were used for their economies and, in particular, complained that their turnover taxes were included in CC estimates of their NNP, while indirect taxes of other countries were excluded. The shift to NNP at market prices from factor cost was made to solve this problem of data comparability.

Third, as early as 1951 but especially from the late 1960s, there was controversy in the General Assembly whenever assessments of some developed countries were reduced while those of some developing countries were increased. At the extreme, the proposal was sometimes advanced that as long as the disparity between developed and developing countries persisted, the total assessment of developed countries should not decrease at the expense of that of developing countries. Counter-arguments (made by developed countries) were that this suggestion was incompatible with the principles of ability to pay and collective financial responsibility, and that assessment was legitimately on an individual-country rather than group basis.

Fourth, the 'middle-income' countries, essentially developed countries other than the United States, view themselves as an unprotected group, enjoying neither the ceiling (or, while in effect, the percapita ceiling, except for a few countries) nor low-per-capita-income allowance and other concessions to the developing countries, and thus most subject to passive increases in assessment.

These tensions suggest that the scale be examined from the standpoint of four country groups: the United States (considered a one-country group), other developed countries, Eastern Europe inclusive of the USSR, and developing countries. Basically, countries are assigned to the groups according to the classification of the United Nations Statistical Office (UNSO) in its various publications.[11]

The inclusion of many scale elements and/or changes in the values of the parameters of the elements were a result of country-group pressure, conflict, and compromise. The first change in the income concept, from factor cost to market prices, can be viewed as a response to Soviet-bloc complaint. The second change, incorporation of debt adjustment, emanated from developing-country pressure. The very existence of the ceiling in the scale and every reduction in its value were results of U.S. initiative and power, albeit the ceiling level was subject to compromise. The developing-country group has fought for reductions in the floor level; naturally, it has been opposed by the developed countries. The per-capita ceiling (now defunct) can be interpreted as an effort by some higher-income countries (especially Canada and Sweden) in the 'middle-income' group to achieve protection from being passive recipients of points resulting from other-groups' influence on the scale. The developing-country and Soviet-bloc groups traditionally were vociferous in their opposition to both ceilings.

While the existence of the low-per-capita-income allowance in the scale was not a result of developing-country pressure, the increase in the values of its parameters over time certainly was. The developing group pushed successfully for other concessions, including 'no increase in the assessments of the least developed countries,' a longer statistical base period, and the scheme of limits (the last two elements considered a means of protecting growing countries of the group from increases in assessment). Much mitigation activity by the CC has been a response to developing-group insistence that increases in its assessments be moderated, that decreases be initiated, and that its problems of external debt, foreign exchange, and natural disasters be recognized. With so many inter-group tensions and conflicts at work, the very principle of limited assessment change can be considered a technique of reducing the impact of these confrontations.

7. Positive analysis of UN Scale

To this point both the official scale and the UN procedure of obtaining the scale have been accepted. Now the scale is retained but the UN procedure is replaced by an alternative approach. The early writers on international burden-sharing described a progressive assessment as 'each country contributing a percentage of its national income but with this percentage dependent on relative income per capita' (Schelling, 1955, p. 12).[12] Though these authors did not suggest it, a direct expression of the relationship in real-income terms follows naturally. The relationship involves progressivity with respect to only per capita income. Suppose that it is complemented by progressivity with respect to total income, with the amount of each type of progressivity left open, and that country-group effects are included. Then one may test the extent to which the UN scale methodology plus ad hoc adjustment – constituting a complex, multi-step procedure, in large part a resultant of country-group conflict and compromise, and based on nominal income – can be replaced by a simple model, paying explicit attention to country groups, and grounded on real income. The relationship, however, cannot be directly applicable to countries assessed at the ceiling or floor, because the country assessment may be set by the constraint rather than by the scale methodology (plus ad hoc adjustment) in the UN or by the model here.

The relationship may be stated as an econometric equation as follows, for a given scale (pertaining, of course, to a given scale period):

$$R_i = K \cdot F^{DC_i} \cdot G^{EE_i} \cdot YK_i^H \cdot u_i \qquad (1)$$

R_i is country i's assessment-income ratio normalized by the mean ratio over the countries in the sample. Specifically, $R_i \equiv (A_i/Y_i)/(mean\,(A_i/Y_i))$, where A_i is country i's official assessment rate and Y_i its real income. The mean R_i over countries in the sample is unity.

DC_i and EE_i are dichotomous variables with value unity if country i is in the developed-country (DC_i) or Eastern European (EE_i) group, respectively, zero otherwise. $YK_i \equiv YP_i/YP_{US}$, the ratio of country i to U.S. per capita real income. Parameters of the equation are K, F, G, H; and u_i is country i's error term, assumed to be identically and independently distributed for all i, with an expected value of unity.[13]

Country-group effects are as follows. The percentage effect on the assessment-income ratio of being developed is $100 \cdot (F-1)$ and that of being Eastern European is $100 \cdot (G-1)$. Favoritism to developing countries is implied by F, $G > 1$. Progressivity of the scale is measured as follows. According as the elasticity (H) of the assessment-income ratio (R_i) with respect to per capita income (YK_i) is greater than, equal to, or less than zero, the scale may be deemed progressive, proportional, or regressive with respect to per capita income. Recalling that R_i is A_i/Y_i normalized to a mean of unity; according as K (for developing countries), $K \cdot F$ (for developed countries), or $K \cdot G$ (for Eastern Europe) is greater than, equal to, or less than unity, the scale (subset of A_i) for that group is progressive, proportional, or regressive with respect to total income (Y_i).

The analysis in this section is positive (because the UN scale, the set A_i, underlies the dependent variable) and disaggregative (in the sense that all countries in the sample are given equal weight in estimating equation (1)). While its official scale is accepted, the UN measure of assessable income, meaning income inclusive of the low-per-capita-income allowance, is not.[14] The variable Y_i differs from UN assessable income in five respects: the income concept is gross national product (**GNP**) rather than **NNP**, in constant rather than current prices, and exclusive of debt adjustment for all scales; the conversion factor is **PPP** rather than **ER**; and progressivity is determined by the data and specified equation rather than predetermined. However, the unweighted average of annual income is taken over the statistical base period, and the mid-period population figure is used to compute per-capita income, both in accordance with UN procedure.[15]

Equation (1) is estimated separately for all scale periods delimited in Section 2 and listed in Table 8.1. This moving cross-section analysis provides answers to four questions. First, how close does the equation approximate the UN scale for the various scale periods? Second, is the UN scale progressive in terms of real income and how has its progressivity changed over UN history? Third, are there country-group effects (for example, favoritism to developing countries) distinguishable from the influence of per-capita income in determination of the scale, and how have these effects changed over time? Fourth, how have the highest-assessed countries fared relative to the equation's predictions?

To estimate equation (1), ordinary least-squares is applied to the equation transformed into logarithmic form.[16] Among the constant term and the variables DC_i, EE_i, and YK_i, the subset that maximizes $\bar{R}^2$ is included in the regression. Because the initial statistical base period was prewar, the CC included effects of 'war dislocation' and 'war improvement' in the preparation of its original recommended scale. Therefore a dichotomous variable, DI_i (1 if dislocated by World War II, 0 otherwise), is included in the equation for the 1947 and 1950 scale periods.[17] The samples exclude the United States (always the sole ceiling country), all countries assessed at the floor, and China until 1980–82.[18]

Estimates of equation (1) are presented in Table 8.3, with t-values in parentheses below coefficients. There are two measures of sample size: N, the number of countries, and ΣA, the total assessment of the countries in the sample.

Table 8.3 Regressions for assessment-income ratio

		Coefficient[b] of			*Goodness of fit*		*Sample size*	
Scale period	*Constant[a]*	*DC*	*EE*	*Income elasticity[c]*	r^2	*U*	*N*	ΣA
1947	1.62	0.70[d]		0.41	.31	.15	19	42.11
	(2.18)	(1.94)		(2.52)				
1950	0.78	1.57	–	–	.29	.17	26	47.49
	(2.67)	(2.92)						
1951	0.79	1.51	–	–	.27	.18	28	51.09
	(2.69)	(2.94)						
1952	0.77	1.59	1.52	–	.18	.14	45	56.55
	(3.49)	(3.50)	(1.76)					
1953	0.76	1.64	1.81	–	.28	.17	44	58.43
	(3.92)	(3.90)	(2.57)					
1954	0.75	1.67	1.95	–	.34	.20	44	60.22
	(4.14)	(4.10)	(2.94)					
1955	0.73	1.74	2.01	–	.38	.24	44	60.69
	(4.41)	(4.41)	(3.06)					
1958	–	1.29	1.53	0.13	.27	.37	60	61.48
		(2.93)	(2.66)	(4.56)				
1959–61	–	1.47	1.69	0.18	.53	.36	58	61.49
		(4.79)	(3.58)	(6.65)				

1962–64	–	1.48	1.94	0.20	.56	.34	60	61.76
		(4.54)	(4.21)	(6.88)				
1966–67	0.76	1.89	2.25	0.12	.55	.36	60	61.37
	(1.78)	(4.57)	(4.50)	(1.54)				
1968–70	0.74	2.01	2.31	0.13	.59	.35	59	61.91
	(1.86)	(4.77)	(4.50)	(1.63)				
1971–73	0.65	2.33	2.45	0.10	.64	.33	57	61.75
	(2.94)	(6.25)	(5.11)	(1.33)				
1974–76	0.62	2.78	3.00	0.21	.79	.28	59	67.96
	(3.50)	(7.96)	(7.07)	(3.00)				
1977	–	2.00	2.58	0.56	.80	.21	57	67.72
		(9.45)	(7.17)	(16.40)				
1978–79	–	2.29	2.75	0.51	.85	.22	73	68.61
		(10.73)	(7.18)	(17.70)				
1980–82	–	2.26	2.40	0.52	.83	.21	74	74.10
		(10.10)	(5.92)	(17.13)				
1983–85	1.21	2.07	2.04	0.69	.82	.23	72	74.06
	(1.62)	(6.05)	(4.48)	(10.50)				
1986–88	1.39	1.87	1.86	0.79	.78	.23	75	74.05
	(2.81)	(5.22)	(3.85)	(11.47)				
1989–91	1.49	1.73	1.70	0.82	.74	.23	73	73.91
	(2.95)	(4.10)	(2.99)	(10.28)				
1992–94	1.68	1.61	1.38	0.87	.79	.26	68	73.52
	(3.56)	(3.46)	(1.73)	(10.44)				

Source: see section 7 of text.

a A dash indicates that the constant is taken as unity, because its inclusion in the regression reduces $\overline{R}^2$.

b A dash indicates that the coefficient is taken as unity, because inclusion of the variable in the regression reduces $\overline{R}^2$.

c A dash indicates that elasticity is taken as zero, because inclusion of the income variable reduces $\overline{R}^2$.

d Variable is DI.

Variables

Assessment-Income Ratio: normalized to unit mean
DC = 1 if developed country, 0 otherwise
EE = 1 if Eastern European country, 0 otherwise
DI = 1 if dislocated by World War II, 0 otherwise
Income: ratio of per-capita income to U.S. per-capita income

Goodness of fit

r^2 = squared correlation between actual and fitted assessment-income ratio
U = Theil's U statistic for actual and fitted assessment

Sample size

N = number of countries
ΣA = total assessment of countries in sample

The latter statistic understates the sample coverage because the United States, the floor countries, and China until 1980–82 are excluded by default. The percentage coverage, meaning ΣA as a percent of the total assessment of the potential sample, is 78 percent in 1947, 88 percent in 1950, 93 percent in 1951, and over 99 percent in the subsequent scale periods.

Approximation to UN scale Two measures of goodness of fit are used. The first is r^2, the squared correlation between R_i and $\hat{R}_i$, which measures the direct predictive ability of equation (1) over the sample period. However, the ultimate interest is A_i, for which the predictive power of equation (1) may be gauged by Theil's inequality coefficient, U, where $U^2 = \Sigma(\hat{A}_i - A_i)^2/\Sigma A_i^2$. A perfect fit implies U = 0, while the naive model $\tilde{A}_i = 0$, all i, yields U = 1. (See note 16 for the computation of $\hat{R}_i$ and $\hat{A}_i$.) Considering the cross-sectional nature of the equation, the goodness-of-fit is impressive for most scale periods. The r^2 is above .50 from 1959–61 onward and U is below .25 for the majority of the scale periods. From 1958 onward, all explanatory variables are present and their coefficients are frequently highly significant. In all cases coefficients have the theoretically correct sign. Also, the magnitudes of the country-group coefficients are as expected: those of DC and EE both above unity in all scales from 1952 (favoritism toward developing countries), that of DI below unity in 1947 (recognition of war damage). It can be concluded that equation (1) constitutes a surprisingly good model of UN scale determination.

Progressivity and country-group effects Per-capita income elasticity ($\hat{H}$) is always positive, implying scale progressivity with respect to per capita income. Though the coefficient is uniformly inelastic, there is an upward trend in its value. The developing countries face a progressive scale with respect to total income for those scales for which the constant term exceeds unity (1947 and 1983–94). In contrast, except for 1947 and 1947–51, respectively, all developed and Eastern European countries encounter a progressive scale (the product of the constant and the respective coefficient above unity). From the late 1960s to the early 1980s the country-group percentage effects against these groups and in favor of developing countries actually exceeds 100 percent.

Highest-assessed countries In every scale period, the countries with the four highest assessments account for more than half, and sometimes almost two-thirds, of total assessments.[19] So it is of interest to explore the extent to which these countries are overassessed or underassessed in light of equation (1). Table 8.4 lists, for each scale period, the four highest-assessed countries together with their pertinent statistics. Consider first the normalized assessment-income ratio (R_i), shown in the first part of the table. For most scale periods, all four countries have values above, sometimes substantially above, the sample mean (unity) – meaning that they were 'overassessed' by this simple measure.

For the 1954 to 1971–73 periods, the USSR assessment-income ratio is the highest among all the sample countries; whereas from 1952 onward the U.S. ratio is the lowest of the four highest-assessed countries. Yet in all scale periods USSR per capita income is lowest and U.S. highest among the highest-assessed countries. So the ceiling was effective in limiting the U.S. assessment-income ratio relative to that of the other high-assessed countries.

Table 8.4 Statistics for countries with four highest assessments

	Normalized assessment-income ratio[a]				*Actual minus fitted assessment*[b]			
Scale period	*U.S.*	*USSR*	*U.K. Japan*[c]	*France Germany*[d]	*U.S.*	*USSR*	*U.K. Japan*[c]	*France Germany*[d]
1947	1.20	0.72	1.13	1.01	–14.13	1.22	0.84	0.69
1950	1.04	0.89	1.43	1.39	–6.57	0.98	1.70	0.77
1951	1.06	0.92	1.45	1.28	–4.89	1.20	1.99	0.42
1952	1.03	1.33	1.44	1.28	–7.31	1.32	1.58	0.22
1953	0.96	1.70	1.43	1.29	–10.08	2.79	1.37	0.20
1954	0.89	1.91	1.36	1.25	–13.36	3.93	0.79	0.02
1955	0.89	2.06	1.27	1.30	–14.32	5.02	–0.06	0.13
1958	1.05	2.16	1.31	1.46	–7.74	6.77	0.59	1.11
1959–61	1.19	2.16	1.53	1.82	–7.53	6.43	0.99	1.88
1962–64	1.28	2.34	1.63	1.76	–4.92	6.50	1.32	1.54
1966–67	1.37	2.38	1.68	1.83	–1.41	6.72	1.37	1.61
1968–70	1.38	2.36	1.62	1.76	–2.62	6.67	0.88	1.26
1971–73	1.41	2.26	1.59	1.80	–2.34	6.28	0.55	1.19
1974–76	1.29	2.22	1.39	1.85	–8.01	5.41	–0.49	1.06
1977	1.26	1.83	1.51	1.97	–14.52	3.59	0.31	1.44
1978–79	1.45	2.16	1.78	2.25	–14.37	4.02	0.42	1.29
1980–82	1.34	1.85	1.75	2.25	–17.21	3.73	0.35	1.45
1983–85	1.30	1.71	1.80	2.26	–23.13	4.24	0.50	1.21
1986–88	1.33	1.65	1.84	2.27	–24.14	4.18	0.47	1.04
1989–91	1.24	1.54	1.77	2.15	–26.79	4.15	0.09	0.87
1992–94	1.25	1.50	1.90	2.14	–29.20	4.80	0.22	1.26

Source:
Normalized assessment-income ratio and fitted assessment: see section 7 of text. Actual assessment: CC Report and FC Report, various sessions, 1946–1991.

a Dependent variable in Table 8.3 regressions.
b Measured in percentage points. Fitted assessment from Table 8.3 regressions.
c U.K. for scale periods 1947 to 1971–73, Japan for scale periods 1974–76 to 1992–94.
d France for scale periods 1947 to 1971–73, Federal Republic of Germany for scale periods 1974–76 to 1989–91, Germany for scale period 1992–94.

Also, the USSR had justification from 1952 (when its assessment increased by 41 percent over 1951, followed by further annual increases of 25 and 15 percent) in complaining of its assessment relative to that of the high-income Western countries.

Consider now the statistic $(A_i - \hat{A}_i)$, the actual minus fitted assessment for country i, shown in the second part of Table 8.4.[20] This statistic measures, in percentage points, the overassessment (if positive) or underassessment (if negative) of country i relative to the estimated equation (1). The United States always benefits from the ceiling, and its underassessment is at a maximum in the 1980–94 scales.[21] In contrast, the other high-assessed countries are, with two minor exceptions, uniformly overassessed for all scale periods.

The overassessment of the USSR (or ex-USSR, for 1992–94) is shown even more clearly, exceeding that of the remaining high-assessment countries in 1947 and then uniformly from the 1953 scale. Incredibly, the USSR over-assessment dwarfs that of Japan, especially in the most recent scales. Looked at another way, Japan's assessment is close to being 'on the regression line,' while the USSR assessment is 'above the line.'

8. Properties of a normative scale

The third viewpoint of the study involves rejection of not only the UN procedure of obtaining its scale but also the official scale itself. An ideal scale is derived and used to evaluate the official scale. Because it is used in a normative analysis, the ideal scale is called the 'normative scale.' In this section properties of the normative scale are established.

Number of assessment principles The multiplicity of assessment principles in the UN 'scale methodology plus ad hoc adjustment' fosters dissension and bargaining as distinct from adherence to principle. The political difficulty of constructing a scale is magnified, because the relative weights of the various principles require determination and this is both the evident initial step in constructing the scale and the obvious initial subject of conflict. Multiple criteria also make the scale procedure complex, and it can be surmised that a complex procedure enhances the scope for country-group conflict. So a single assessment principle has the dual advantage of defusing political tensions and of simplicity, with the added benefit of a favorable interaction of the two benefits.

Selection of assessment principle Given that a single assessment criterion is to be selected, ability to pay is the obvious choice. It is the most notable principle of national tax systems and is consistent with economists' concepts of horizontal and vertical equity. It also happens to be the most important principle underlying the UN scale procedure both legally and empirically (as shown in Sections 3 and 5, respectively).

Income concept To measure ability to pay, a strong case can be made for a multidimensional approach, inclusive of wealth and of social and economic indicators of development, as has often been discussed in the UN itself. However, data limitations and the objective of simplicity dictate resort to income, here as in the UN. National rather than domestic product and market prices rather than factor cost (the first always UN practice, the second since the 1965 scale) are the better indicators of ability to pay (for all scale periods). However, contrary to UN procedure, gross rather than net product (**GNP** rather than **NNP**) is taken. The use of **GNP** enhances international comparability, because accounting depreciation varies with tax law and rulings and with data quality. Also, with **GNP** measuring total output, it is a matter of choice how much of it is devoted to depreciation – and it is arguable that ability to pay be independent of that choice.

The UN shift to debt-adjusted income is not followed. This inclusion of a capital-account transaction in current income violates a national-accounting precept. Also, asymmetrically, while subtracting repayment of debt principal from income, the UN concept adds neither the original borrowing to debtor's income nor the repaid principal to lender's income.

Denomination of income **GNP** in constant rather than current prices is adopted (contrary to UN practice), because mere price change does not alter a country's ability to pay. To convert constant-price **GNP** to a common currency, **PPP** is utilized, in opposition to the CC selection of **ER** supplemented by **PARE**. The advantages of **PPP** over **ER** are threefold. First, **ER** is based on a weighting pattern of 'output' that is nebulous and related directly only to tradables; whereas **PPP** (**GNP** concept) has a precise weighting pattern based on shares of tradables and nontradables (and their components) in national production. Second, in contrast to **ER**, **PPP** is not subject to the **ERCB**. Third, **ER**, unlike **PPP**, is influenced by factors irrelevant to international income comparison, for example, speculation and exchange-market intervention. The disadvantages of **PARE** relative to **PPP** are that (i) the former corrects only for differential changes in price levels relative to exchange rates rather than providing appropriate conversion factors for price levels themselves, and (ii) the amount of **PARE** correction varies with the base year chosen and in fact can be negative.[22]

Progressivity The amount of progressivity that the scale should contain is inherently arbitrary. Therefore the UN explicit progressivity is accepted, which enhances comparability with the official scale. Specifically, the UN low-per-capita-allowance formula is taken, but with the per-capita-income limit re-expressed in real-income terms. Note that the elements in the UN scale under the ability-to-pay criterion that affect progressivity implicitly (**ER** conversion factor, debt adjustment) are excluded, as are, of course, all elements under the other assessment principles and the entirety of ad hoc adjustment.

Base period A three-year 'normative base period' is selected, because it is a good compromise between the principal advantage of a shorter base period (reflecting current economic conditions) and that of a longer one (reducing the effect of fluctuations in income). The early statistical base periods of less than three years are retained as evolutionary to the three-year period; but the CC abandonment of the three year-period in 1977 is rejected because it was founded on the limited assessment change principle rather than ability to pay.[23]

Attention to country groups Because the UN scale is so reflective of country-group tensions, the 'normative scale' should be oriented to an evaluation of the assessments of country groups. In contrast, the UN procedure for setting the scale is on an individual-country basis.

9. Normative analysis of UN scale

Application of the methodology of Section 8 yields (for each scale period) a normative assessment, N^j, for each group j, where j = United States (US), other developed countries (DC), Eastern Europe (EE), and developing countries (DL).[24] The normative assessment can be compared with the actual assessment, A^j, for each country group.[25] The actual assessment (A^j) and actual-normative assessment differential ($A^j - N^j$) for each country group are shown in Table 8.5 for each scale period. Looking at actual assessment (A^j),

Table 8.5 Actual and normative assessments, by country group (percentage points)

	United States		*Other developed countries*		*Eastern Europe*		*Developing countries*	
Scale period	*Actual*	*Actual minus normative*	*Actual*	*Actual minus normative*	*Actual*	*Actual minus normative*	*Actual*	*Actual minus normative*
1947	39.89	2.25	30.92	0.89	9.58	–2.95	19.61	–0.20
1950	39.79	–9.40	30.44	7.83	9.58	1.31	20.19	0.26
1951	38.92	–8.89	30.24	6.82	10.18	1.77	20.66	0.31
1952	36.90	–14.36	28.74	4.32	13.90	4.82	20.46	5.23
1953	35.12	–16.88	28.24	2.09	16.97	8.79	19.67	6.00
1954	33.33	–18.58	27.73	1.61	19.31	11.07	19.63	5.90
1955	33.33	–18.67	27.21	1.33	20.28	11.99	19.18	5.35
1958	32.51	–14.81	30.36	–1.06	19.29	11.12	17.84	4.75
1959–61	32.51	–10.63	31.70	–1.93	19.09	9.68	16.70	2.87
1962–64	32.02	–8.87	30.59	–2.93	21.04	10.76	16.35	1.05
1966–67	31.91	–4.84	31.18	–3.41	21.09	9.49	15.64	–1.24
1968–70	31.57	–2.88	32.31	–3.17	20.54	8.73	15.58	–2.67
1971–73	31.52	–1.26	33.51	–0.95	19.92	6.09	15.05	–3.89
1974–76	25.00	–4.42	41.95	1.92	19.30	7.48	13.75	–4.98
1977	25.00	–3.24	43.30	3.39	17.60	5.37	14.10	–5.53
1978–79	25.00	–2.60	43.75	4.03	17.82	5.46	13.43	–6.89
1980–82	25.00	–0.82	47.19	9.63	17.12	3.23	10.69	–12.04
1983–85	25.00	–1.20	48.99	11.54	15.70	2.32	10.31	–12.67
1986–88	25.00	–0.31	49.33	12.46	15.07	2.03	10.60	–14.18
1989–91	25.00	–0.09	49.45	14.76	14.63	2.23	10.92	–16.90
1992–94	25.00	0.66	51.20	16.96	12.41	1.61	11.41	–19.23

Source:
Actual assessment: CC Report and FC Report, 1st – 46th sessions, 1946–1991. Normative assessment: see Appendix A.

other developed countries increase their assessment almost throughout UN history (except for the early 1950s), so that by 1992–94 they account for over half the total assessment. In contrast, the United States to 1974–76, Eastern Europe from 1968–70, and developing countries from 1952 enjoy almost continuous reductions in their assessment.

Group overassessment ($A^j - N^j > 0$) or underassessment ($A^j - N^j < 0$) follows patterns. The effect of the ceiling on the U.S. assessment is strongest in 1950–64, with the constraint becoming inconsequential and eventually nonbinding in 1980–94.[26] Other developed countries are overassessed in the early scales, underassessed in 1958–73, and then increasingly overassessed. After 1947, Eastern Europe is always overassessed; but there is a downward trend from 1962–64. Developing countries are overassessed in 1950–64; but from 1953 there is a steady decrease in their overassessment, and there is a switch in 1966–67 to steadily increasing underassessment. While *collective financial responsibility* (represented by the ceiling) has had reduced impact over time, *level of development* has become increasingly powerful.

The total intergroup transfer of points, $(\Sigma \mid A^j - N^j \mid)/2$, measures how close the official scale is to its normative equivalent. Of the 21 scale periods, nine involve a 'fit' of over 90 percent (corresponding to an intergroup transfer of under ten percentage points) and twelve a 'fit' of 80–90 percent.

10. Summary and conclusions

Three approaches are used to examine the UN scale of assessments, meaning the apportionment of UN expenses among its member countries, over the entire UN history. First, *both the scale and the UN procedure of determining it are accepted*. It is shown that the elements in determination of the scale emanate from various assessment principles, that the UN use of the exchange rate to express income in a common currency involves a hidden progressivity in the scale, and that the relative importance of the various elements in the scale is measurable (though information is complete only for recent scales).

Second, *the UN scale is accepted but the procedure is simplified and rationalized*. Progressivity is found to increase over time and there are strong country-group effects in favor of developing countries. Regarding the highest-assessed countries, the United States is underassessed, the USSR overassessed, and Japan assessed close to what the model predicts.

Third, *the UN scale is rejected in favor of a normative alternative*. From this standpoint, the U.S. underassessment is small by the 1980s, while developing countries enjoy increasing underassessment from the mid-1960s. Overassessment of Eastern Europe tends to decrease from the mid-1950s, but that of developed countries other than the United States increases from the mid-1970s. Due largely to the overassessment of other developed countries, by 1992–94 developed countries (including the United States) accounted for over three-quarters of the UN scale. In contrast, and in spite of strong and persistent complaints about their apportionment of UN expenses, the developing countries by 1992–94 paid almost twenty percentage points less than their normative share.

Appendix A: Computation of normative scale

Step 1: Compute **PPP**-converted **GNP** in constant prices ('real **GNP**,' Y_i) and per-capita real **GNP** (YP_i) for each country i for which data are available, for the given normative base period (with the set of resulting i denoted as s2). In accordance with UN practice, (i) the *unweighted average* of annual **GNP** and (ii) the *mid-period* population, are taken from the $GNPM_i$, and P_i series, respectively, in step 7 of Appendix B.

Step 2: Convert the per-capita-income limit of the UN low-per-capita-allowance formula to its real-**GNP** equivalent. This is done by multiplying the UN limit for the base period by the ratio of 'U.S. real **GNP** (Y_{US})' to 'U.S. nominal **NNP** (national income, for 1947 to 1957–59) averaged over the base period.'[27]

Step 3: For each country calculate *real assessable income* (YB_i), that is, real **GNP** adjusted for the low-per-capita-income allowance. The converted income limit of step 2 is applied with the UN gradient.

Step 4: Sum the YB_i into country groups: $YB^j = \Sigma YB_i$, $i \in j$, where j = US, DC, EE, DL.

Step 5: Adjust the YB^j for differential sample coverage: $YC^j = (A^j/A_s^j)\cdot YB^j$; where $A_s^j = \Sigma A_i$, $i \in j$ and $i \in s2$, is the sample coverage for group j (see note 25).

Step 6: Adjust the YC^j so they sum to total assessments (given by the last column of Table 8.1 for the specific scale period). The result is the normative scale: $N^j = (\Sigma A^j/\Sigma YC^j).\ YC^j$

Step 7: Compute the sample coverage, ΣA_s^j. For 1947, 1950, and 1951, coverage is 82, 87, and 90 percentage points. For the remaining scale periods, it exceeds 99 percent.

Appendix B: Income and population data

The objective is to have, for each country i, an annual time series of (1) **PPP**-converted, U.S.-dollar-denominated **GNP** in constant prices and (2) population, enveloping the statistical (and hence, by default, normative) base periods corresponding to the scale periods for which country i was a UN member. Further, these time series should be consistent both over time and across countries. The series are obtained from the following steps.

Step 1: From Summers and Heston (1991), obtain the annual series on real **GDP** per capita (international dollars, 1985 prices, Laspeyres index), denoted as $GDPL_i$, and on population (thousands of persons), denoted as P_i, for each UN member i for as much of the requisite time period as their series encompass.[28] Their maximum coverage for a given country is 1950–88.[29] These Summers–Heston **GDP** data, except for their divergence from **GNP**, fulfill the conceptual requirements to serve as the required input for step 1 of Appendix A: the year 1985 emanates from a **PPP** computation; figures for other years are extrapolated using national-accounts series at constant prices (Summers and Heston, 1991, p. 343).

Step 2: Compute real **GDP** in millions of international dollars: $GDPM_i = (P_i \cdot GDPL_i)/1000$.

Step 3: Multiply $GDPM_i$ by the ratio of **GNP** in current prices to **GDP** in current prices (for the years for which data on this ratio exist–otherwise assume a unity ratio).[30] The result is $GNPM_i$, real **GNP** in millions of international dollars.

Step 4: For cross-country consistency, the Summers and Heston (1991) data fix the country coverage. However, it is possible, and indeed desirable, to extend their temporal coverage for the given countries. Let $GNPM_i^l$ denote $GNPM_i$ for the latest available year, l (usually, though not always, 1988), and $GNPM_i^e$ for the earliest available year, e (1950 at the earliest). Using all available published data sources, two series of output at constant domestic prices, denoted as $GNPC_i$, are developed, data permitting.[31] One series begins in year l and ends in 1989; another begins in 1938–40 (the average of whatever of these years are available) and ends in year e.[32] The first series extends $GNPM_i$ forward to 1989 (or as close to that year as data availability permits) by means of the multiplicative factor $GNPM_i^l/GNPC_i^l$; the second series extends $GNPM_i$ backward to 1938–40 (or 1950, or as far back as data permit) by means of the multiplicative factor $GNPM_i^e/GNPC_i^e$, where $GNPC_i^l(GNPC_i^e)$ is $GNPC_i$ in year l (e).[33]

Step 5: Using the same technique as for $GNPM_i$, extend P_i forward to 1991 and backward to 1938–40 (or rather as much in either direction as correspondence with the extended $GNPM_i$ series warrants).[34]

Step 6: Adjust the $GNPM_i$ and P_i series for merger or separation of countries. For Germany for 1980–89, add the $GNPM_i$ (and P_i) series for the Federal Republic of Germany and the German Democratic Republic. For the United Arab Republic for 1955–57, add the series for Egypt and Syria. For Pakistan (East plus West) for 1959–68, add the $GNPM_i$ series for Bangladesh and Pakistan. Extend the series backward to 1949 via $GNPC_i$ for Pakistan (East plus West). Similarly, for the 1950–68 P_i, add the series for Bangladesh and Pakistan, and extend it to 1949 via a population series for Pakistan (East plus West).

Step 7: The resulting annual $GNPM_i$ and P_i series are the inputs into (i) step 1 of Appendix A, followed by the remainder of Appendix A, for Section 9 and (ii) step 1 of Appendix A revised so that 'statistical base period' replaces 'normative base period,' for Section 7.

Notes

* Reprinted from *International Comparisons of Prices, Output and Productivity*, vol. 1, D. S. Prasada Rao and Jorge Salazar-Carrillo (eds.), pp. 143–176. Copyright 1996, with permission from Elsevier.

1 For a related study, but focused on the developing countries and the time period 1986–94, see Officer (1994). Sources of institutional information about the scale of assessments are United Nations, General Assembly, Report of the Committee on Contributions (cited as CC Report) and Scale of Assessments for the Apportionment of the Expenses of the United Nations: Report of the Fifth Committee (cited as FC Report), 1st–46th sessions, 1946–1991, and Note by the Secretariat, Evolution of the Methodology for the Scale of Assessments and its Current Application (cited as Evolution), 24 April 1989; United Nations, Yearbook of the United Nations, 1946–47 to 1986; U.S. Department of State, United States Participation in the United Nations, various issues 1946–47 to 1989.

2 The regular budget of the United Nations is distinguished from financing of peacekeeping operations. The scale of assessments has always had direct and precise application to the regular budget. Over the UN history, some peacekeeping activities have been paid from the regular budget, but these operations have uniformly been minor in expense and only two (Jerusalem and India-Pakistan) remain. In two other cases (West New Guinea and Yemen), in the early 1960s, parties to the dispute divided the total costs of the UN presence among themselves. One peacekeeping force (Cyprus) has always been financed by voluntary contributions. The original UN Emergency Force (1956–67), stationed in Egypt, and the UN Congo Force (1960–64) were the first truly expensive peacekeeping undertakings. They were financed on the basis of the scale of assessments for the regular budget but as modified by voluntary contributions and rebates to developing countries.

Since 1973, a special assessment scale has been applied to all new peacekeeping operations. Member countries are divided into four mutually exclusive groups: developed countries, paying their respective regular-budget assessment rates; developing countries, paying 20 percent of their regular rates; least developed countries, paying 10 percent; and permanent members of the Security Council, paying their regular rates plus the shortfall from the developing-country groups in proportion to their regular rates. For histories of the financing of UN peacekeeping operations, see Stoessinger (1964) and James (1989).

3 It may be noted that the UN reserves the term ‘element’ only for what are called here the ‘objective elements.’

4 Also, in the 1986–88 scale, World Bank data (that employ a kind of relative-**PPP** concept) were used for three countries. For some scales in the 1970s, the problem of conversion was obviated for a few countries, by direct estimation of **NNP** in U.S. dollars.

5 The implicit real progressivity of a nominally proportional tax was recognized by writers in the traditional burden-sharing literature, though they did not distinguish the two senses of progressivity. See Neale (1961, pp. 36–37), Uri (1963, pp. 46–47), and Pincus (1965, pp. 66–67).

6 For a complete description of the scale methodology for the 1986–94 scales, see CC Report, 44th sess., 1989, pp. 5, 8, and 46th sess., 1991, pp. 7–9.

7 Only in the year 1959 did the CC agree to permit a member state to see its own specific data (income, conversion factor, population, low per-capita-income allowance) – and that in response to a GA resolution (December 10, 1958). It refused to release data for other countries to an individual member, even though, obviously, ‘it was difficult to ascertain whether increases in assessments were justified without access to the national income data for all Member States’ (FC Report, 25th sess.,

1970, p. 2). Member countries were treated no differently from scholars; they had to wait patiently for the CC to publish in its future Reports such parts of the data documentation underlying its work as it decided. In fact, the CC released neither the machine scale nor country-comparative income data until the 1983–85 scale.

8 This element of the scale methodology redistributed 0.02 percentage points in 1983–85 (CC Report, 37th sess., 1982, p. 9) and zero points in the later scales.

9 Data sources are CC Report, 38th sess., 1983, pp. 24–25; 45th sess., 1990, pp. 45–56; 46th sess., 1991, pp. 29–36.

10 It may be noted that $\Sigma A_i = 100.00$ while $\Sigma^* A_i = 98.78$, the latter the percentage sample coverage. For the other scale periods, the sample coverage is 100 percent, except that two new members (Marshall Islands and Micronesia, with assessment 0.01 each) are excluded from the 1992–94 computation.

11 'Other developed countries' consists of the members of the Organization of Economic Co-operation and Development, but excluding the United States and Turkey, plus Liechtenstein, South Africa, Malta, and Israel, the last two countries beginning with the 1974–76 scale (consistent with the 1973 decision period and the UNSO reclassification of these countries from developing to developed status in 1972–73). 'Eastern Europe' comprises the countries of the traditional Soviet bloc: Albania, Bulgaria, Czechoslovakia, German Democratic Republic, Hungary, Poland, Romania, the USSR or the countries of the ex-USSR (1992–94 scale), and Yugoslavia (1946–50 scales). The UNSO reclassification of Turkey and Yugoslavia from developed to developing (in 1972 and 1985, respectively) is an illogical direction of change; these countries are considered always developing (except Yugoslavia when in the Soviet bloc).

12 Similar or related statements are made by Robbins (1950, p. 16), Hoag (1957, p. 530), Neale (1961, p. 36), and Uri (1963, p. 46).

13 Homoscedasticity (the key to an identical distribution across i) is a reasonable assumption, because the dependent variable ($R_i \equiv normalized\ A_i / Y_i$) is in ratio form. (It would be decidedly unreasonable were A_i the dependent variable and Y_i an explanatory variable.) Because the data are cross-sectional, independence of u_i from u_j, $i \neq j$, can be accepted. Independence of u_i, all i, from DC_i and EE_i is assured by the variables' nonstochastic status. While YK_i is stochastic, it may reasonably be assumed to be distributed independently of u_i, all i. The reason is that a country's UN dollar assessment – one-hundredth the UN budget multiplied by A_i or, equivalently, by $R_i \cdot Y_i \cdot mean(A_i / Y_i)$ – is extremely small relative to the country's income. For example, excluding floor countries, the ratio of 'dollar assessment for 1989' to 'nominal NNP averaged annually over the statistical base period, 1977–86,' ranges from 0.002 to 0.012 of one percent (CC Report, 44th sess., 1989, pp. 42–46). These numbers are too low for a discernible effect of a country's assessment rate (A_i) on its income (Y_i) and thence on per capita income (YP_i).

14 Under the allowance, the relief (income subtracted from low-income countries) is redistributed to high-income countries (those with per-capita income above the limit) in proportion to their incomes. Prior to the 1980–82 scale, the relief was distributed to all countries, both low-income and high-income, which reduced the progressive effect of the allowance.

15 Mid-period population means the middle year of a base period with an odd-year length, the average of the two middle years for a base period with an even-year length. Justification of the income concept and conversion factor is provided in Section 8. Data sources and construction of variables are in Appendix B and the first paragraph of Appendix A.

16 The fitted regression is $ln\hat{R}_i = \hat{k} + \hat{f} \cdot DC_i + \hat{g} \cdot EE_i + \hat{h} \cdot lnYK_i$, where $\hat{k}$, $\hat{f}$, $\hat{g}$, $\hat{h}$ are least-squares estimates of its parameters and $ln\hat{R}_i$ is the fitted value of lnR_i. Estimates of the parameters of equation (1) are $\hat{K} = e^{\hat{k}}$, $\hat{F} = e^{\hat{f}}$, $\hat{G} = e^{\hat{g}}$, $\hat{H} = \hat{h}$; the fitted value of R_i is $\hat{R}_i = e^{ln\hat{R}_i}$; and the fitted value of A_i is $\hat{A}_i = \hat{R}_i \cdot Y_i \cdot mean\ (A_i / Y_i)$.

17 It is not clear whether the CC procedure took the form of updating income figures or of ad hoc adjustment, but no matter. While the GA altered the CC initial scale, it did so principally by imposing a ceiling and floor, and the countries subject to these constraints are excluded from the sample. Countries with unity value of DI_i are those that had been occupied (totally or partially) by Germany or Japan, plus the United Kingdom. DI_i is applied to the 1950 scale period because of minimal changes in the scale through that period.

18 Inclusion of the United States and the floor countries could seriously bias the regression, as extreme observations (the one having an assessment-income ratio below, the others above that predicted by its per-capita income and country group). China's assessment was far above that justified by its income when the country was represented by the Republic of China, because of the legal fiction that the Republic represented all of China and the initial establishment of China's assessment at the level of France, for lack of Chinese data. The gross overassessment continued after representation was switched to the People's Republic, until China submitted income and population data (1980–82 scale). The inclusion of China in earlier samples would severely bias the regression in the same direction as the floor countries.

19 China, tied with France for the fourth highest assessment in 1947–52, is excluded from the analysis.

20 Although the United States is not in the sample, its $\hat{A}_i$ may be computed as for other countries.

21 The effects of the ceiling in Table 8.4 differ from those in Table 8.2 because the point of reference is the estimated equation (1) as distinct from the UN scale methodology.

22 An important reason given by the CC for its rejections of constant-price data and of **PPP** is lack of data; but it can be argued with some justification that this excuse no longer has validity (see Appendix B).

23 The base period was extended 'in order to alleviate the sharp variations in the rates of assessment of countries whose national incomes had risen rapidly in the early 1970s' (CC Report, 42nd sess., 1987, p. 9). The normative base periods are presented in Table 8.1.

24 Derivation of the normative scale is provided in Appendix A.

25 Note that $A^j = \Sigma A_i$, $i \in j$, and that $\Sigma A^j = \Sigma N^j$ = total assessments, given in the last column of Table 8.1.

26 These results are strikingly divergent from those of Section 7 and Table 8.4; but the 'norm' there (a fitted regression equation) is substantially different.

27 The source of U.S. nominal data is U.S. Department of Commerce (1986) and Survey of Current Business, various issues.

28 The population figures for Nigeria in 1957–59 are corrected by adding 10 million.

29 For countries in their Table B3, GDPL for 1985 is computed as the product of .01, relative per capita **GDP**, and U.S. 1985 GDPL.

30 Data sources for the ratio are World Bank, World Tables, and UN, Yearbook of National Accounts Statistics, various issues.

31 The data sources fall into six categories: (1) earlier versions of the Penn World Table [of which Summers and Heston (1991) constitute the fifth version]: Summers, Kravis, and Heston (1980), Summers and Heston (1988); (2) World Bank, World Tables, various issues; (3) historical-statistics volumes: Mitchell (1975, 1982, 1983), U.S. Bureau of the Census (1975); (4) United Nations, Statistics of National Income and Expenditure and Yearbook of National Accounts Statistics, various issues; (5) publications of the Research Project on National Income in East Central Europe: Alton (1980), Alton and others (1990), Lazarcik (1969); (6) Central Intelligence Agency, Handbook of Economic Statistics, various issues.

32 Different series of $GNPC_i$ or the same series with different base years are

linked on the basis of one-year overlaps. Should a consistent series have gaps, it is interpolated linearly to obtain the missing annual observations.

33 This aggregative extrapolation technique is consistent with the Summers – Heston Laspeyres income concept while enabling expansion of the observations beyond that limited by their approach of extrapolating on a disaggregative basis (consumption, investment, government spending, exports, and imports separately).

34 Data sources are UN, Demographic Yearbook, and World Bank, World Tables, various issues.

References

Alton, Thad P. (1980), *Economic Growth in Eastern Europe 1965–1979*. Occasional Paper No. 59, Research Project on National Income in East Central Europe. New York: L.W. International Financial Research.

Alton, Thad P., and others (1990), *Economic Growth in Eastern Europe 1975–1989*. Occasional Paper No. 110, Research Project on National Income in East Central Europe. New York: L.W. International Financial Research.

Hoag, Malcolm W. (1957), 'Economic Problems of Alliance,' *Journal of Political Economy*, **65**, pp. 522–34.

James, Alan. (1989), 'The Security Council: Paying for Peacekeeping,' in David P. Forsythe, *The United Nations in the World Political Economy*, pp. 13–35. New York: St. Martin's Press.

Lazarcik, Gregor. (1969), *Czechoslovak Gross National Product by Sector of Origin and by Final Use, 1937 and 1948–1965*, Occasional Paper No. 26, Research Project on National Income in East Central Europe. New York: Columbia University.

Mitchell, B. R. (1975), *European Historical Statistics, 1750–1970*. New York: Columbia University Press.

Mitchell, B. R. (1982), *International Historical Statistics: Africa and Asia*. New York: New York University Press.

Mitchell, B. R. (1983), *International Historical Statistics: The Americas and Australasia*. Detroit: Gale Research Company.

Neale, Alan D. (1961), *The Flow of Resources from Rich to Poor*. Occasional Papers in International Affairs, No. 2. Harvard University, Center for International Affairs.

Officer, Lawrence H. (1994), 'An Assessment of the United Nations Scale of Assessments from a Developing-Country Standpoint,' *Journal of International Money and Finance*, **13**, pp. 415–28.

Pincus, John. (1965), *Economic Aid and International Cost Sharing*. Baltimore: Johns Hopkins Press.

Robbins, Lionel. (1950), 'Towards the Atlantic Community,' *Lloyds Bank Review*, pp. 1–24.

Schelling, Thomas C. (1955), *International Cost-Sharing Arrangements*. Essays in International Finance, No. 24. International Finance Section, Princeton University.

Stoessinger, John G. (1964), *Financing the United Nations System*. Washington, DC: The Brookings Institution.

Summers, Robert, and Alan Heston. (1988), 'A New Set of International Comparisons of Real Product and Price Levels Estimates for 130 Countries, 1950–1985,' *Review of Income and Wealth*, **34**, pp. 1–25.

Summers, Robert, and Alan Heston. (1991), 'The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950–1988,' *Quarterly Journal of Economics*, **106**, pp. 327–68.

Summers, Robert, Irving B. Kravis, and Alan Heston. (1980), 'International Comparisons of Real Product and its Composition, 1950–77,' *Review of Income and Wealth*, **26**, pp. 19–66.

Uri, Pierre. (1963), *Partnership for Progress*. New York: Harper & Row.

U.S. Bureau of the Census. (1975), *Historical Statistics of the United States, Colonial Times to 1970*. Washington, DC: U.S. Government Printing Office.

U.S. Department of Commerce. (1986), *The National Income and Product Accounts of the United States, 1929–82: Statistical Tables*. Washington, DC: U.S. Government Printing Office.

9 Afterword to Part II

In this chapter, unlike in Chapters 6–8 and unlike in the existing literature, financing of the International Monetary Fund (IMF or Fund) and the United Nations (UN) is considered together. This approach both facilitates comparisons and allows an integrated approach to financing of these organizations (the topic of 'Net Financing Advantage' below).

General characteristics of financing the IMF and UN

The UN scale of assessments is directly determined as the percentage allocation of UN expenses among member countries. Thus, normally (that is, apart from anomalies due to change in membership between scales), the total of assessments is 100 (percent). The IMF directly determines quotas denominated in Special Drawing Rights (SDRs); the country percentage distribution of quotas is thereupon derived, readily enough, by arithmetic.

The UN considers overall assessments more frequently than the IMF reviews quotas. Between 1946, when the initial scale of assessments was decided, and 2003, when the scale for 2004–2006 was determined, the UN determined the overall scale a total of 29 times (Table 8.1, notes; United Nations 2003: 46–48). In contrast, from 1944, the year of the Bretton Woods Conference, and 2003, the year of completion of the Twelfth General Review, the IMF formally reviewed overall quotas a total of only 14 times – Bretton Woods, four 'quinquennial' reviews, eight 'general' reviews, and one review outside the normal five-year cycle (International Monetary Fund 2003: 16).

Both the IMF and UN country percentage distribution of financing have a tendency to be resistant to substantial change. There is inertia in the UN scale and the IMF percent distribution of quotas; but, with minor exception, the reasons differ. Consider the IMF first. Of the 13 reviews of overall quotas that followed the initial distribution decided at Bretton Woods, five involved no quota changes at all! This is not surprising, because a sure way to preserve a member-country's percentage share of quotas is to vote against *any* quota increase at a general review – and an 85 percent majority of the total voting power in the Fund is required for any quota increase. This large majority provides a mechanism for countries with a large current quota share but a declining quota share according to formulas to block a general quota increase unilaterally (the United States) or via coalescing with like-minded members with substantial quotas and therefore substantial voting power (other countries of large economic size). As James M. Boughton, official Historian of the Fund, states: 'Winning approval for general quota increases has always been a politically contentious task' (Boughton 2001: 854).

What of the seven reviews that did result in a general increase of quotas? When the IMF has a general review of quotas, it is only the *increase* in quotas (in SDRs) that is to be decided; the quota base (set of existing country quotas) is taken as given. That in itself provides a strong inertia in the existing quota distribution. Then the quota increase is divided into a part proportionate to existing quotas (which obviously does not change the existing percentage distribution of quotas) and a non-proportionate part ('selective and ad hoc elements' – largely determined by quota formula, and which does change the distribution). Of the eight quota reviews that followed Bretton Woods, seven enable the distinction between proportionate and non-proportionate quota increases. Of these seven, six involve a greater proportionate than non-proportionate increase (see International Monetary Fund 2003: 16).

Three other elements change the percentage distribution of quotas, but their quantitative significance is small: ad hoc quota increases, the small-quota policy and quota increases in connection with compensatory financing of export fluctuations, and assignment of quotas to new members. In the first sixty years of the Fund, 1946 to 2005, there were only 19 cases of ad hoc increases in quota outside the context of general reviews and only nine within the context of general reviews (both listed in International Monetary Fund 2005: 15). 'Ad hoc quota increases have been used sparingly also outside the context of General Reviews. Since 1970, there have been only four stand-alone ad hoc quota increases' (International Monetary Fund 2003: 14). For the rationales of the ad hoc increases in quota, see International Monetary Fund (2000a: 29, 2003: 27–8, 2005: 15).

The small-quota policy and quota increases in connection with the compensatory-financing facility, discussed in Chapter 6 (note 29), existed only in the 1950s–1960s and 1960s alone, respectively, and their quantitative impact was small. Of course, allocation of quotas to new members inherently reduces relative quotas of existing members – but by the end of the 21st century the Fund had achieved virtually universal membership; so this issue is essentially defunct.

Turning to the UN, a variety of elements in the machine scale makes for inertia in the scale. First, some early scales (1948–1950 1958, and 1966–1967) differed minimally from the previous scale, allowing only for inclusion of new members (see Table 8.1). This corresponds to Fund quota reviews that leave quotas unchanged. Second, several elements limited scale-to-scale changes in assessments. The scheme of limits, in existence for the 1986–1994 scales and phased-out in the 1995–2000 scales, set a maximum to both absolute-percentage and percentage-point changes in assessments (see United Nations

2003: 47). Some related predecessor elements had existed in the early 1950s (see Chapter 8, Section 4). Also, for the 1983–1997 scales, the assessments of 'least developed countries' were not to increase (see United Nations 2003: 47).

Third, a number of assessment ceilings have existed in UN history. An unqualified ceiling, reduced gradually over time from 39.89 percent in the 1946 scale to 22 percent in the 2001–2003 scale (United Nations 2003: 46–7), has been applicable only to the United States. A per capita assessment ceiling (set at the per capita assessment of the highest assessment country, that is, the United States) ended with the 1976 scale. Beginning with the 1998–2000 scale, 'no increase for least developed countries' was replaced by a ceiling of 0.01 percent in assessment for these countries.

Fourth, there has always been an assessment floor, applicable to a large part of UN membership. Like the unqualified ceiling, the floor has been reduced over time – from 0.04 percent in the 1946 scale to 0.001 percent in the 1998–2000 scale (United Nations 2003: 46–47). The unqualified ceiling affects only one country, but its effect on the scale is substantial. In contrast, the floor affects many countries, but these countries are so small that the total number of percentage points redistributed is unimportant (see, for example, Table 8.2).

Country quotas have four functions in the Fund. First, quotas determine member subscriptions to the organization, which provide the Fund with its lending resources. This role of quotas constitutes 'financing' of the Fund, analogous to financing of the UN via the scale of assessments. Second, a member country's drawings from the Fund are subject to limits based on percentages of the member's quota. Third, voting power in the Fund is largely proportionate to quota. Fourth, SDRs are allocated strictly in proportion to quota. For discussion of the four functions of quotas, see International Monetary Fund (2000a: 8–13; 2000d: 8–9).

These functions are not as inelastic or as rigidly related to quota as a cursory description suggests. The Fund can borrow to obtain additional resources; waivers can override quota-defined limits to drawings, and special lending facilities are not even subject to such limits; 'basic votes' constitute a component of voting power that is equal across countries and independent of quota (although basic votes have been a small and declining component of voting power since 1958 – see International Monetary Fund 2003: 19; 2005: 18); and SDRs have been allocated rarely, in the history of the Fund (though always in strict proportion to quotas).

Several observers have commented that the existence of only one policy instrument (size of quota) to determine four objectives violates the 'Tinbergen rule,' that the number of independent instruments must be at least equal to the number of objectives. Mikesell (1994, pp. 37–38) notes that the possibility of multiple quotas, each based on different variables oriented to the purpose at hand, was discussed at the Bretton Woods Conference. Jha and Saggar (2000: 581–82) suggest 'four different quotas: one for each of the four objectives' and each based on variables specific to the objective, but acknowledge that this is politically infeasible. Kelkar, Yadav, and Chaudhry (2004: 737) also address the situation of 'only one instrument to achieve multiple objectives.' Rapkin and Strand (2005: 1995) make the same observation.

Turning to the UN, its scale of assessments has had a dual role. The scale directly allocates member countries' shares of the regular budget; also, the formula for the 'peacekeeping scale' is derived from the 'regular scale' (see Chapter 7, note 1 and Chapter 8, note 2). It may be that, here also, one policy instrument for more than one policy goal leads to suboptimal results for the objectives.

Finally, what is the subjective component in IMF-quota or UN-scale determination? Quota formulas played only a limited role in quota determination in the early reviews of quotas, but serve as the allocator of the selective (nonproportional-to-quota) component of quotas in quota reviews from the 1980s onward. For details, see International Monetary Fund (2000a: 31–2; 2001b: 6). So, defining the subjective element as the residual determinant of quotas after equiproportional and formula-driven allocations, that element has essentially vanished since the 1980s.

In the UN, the Committee on Contributions or the General Assembly on its own initiative has on occasion used its judgment to reduce assessments of specific countries, which is termed 'mitigation.' Also, ad hoc adjustments (after the scale is approved by the General Assembly) can be made at the request of individual member countries. These elements have always been of minor importance. Furthermore, since the 1992–1994 scale, mitigation is dependent entirely on the willingness of 'donor' countries to accept increases in their assessments to compensate for the mitigation (decreases in assessments of favoured countries). For details on subjective elements in the UN scale, see Chapter 8 (Section 4 and Table 8.2) and United Nations Secretariat (1989: 14–15; 2003: 44–45).

Actual and suggested elements in financing formulas

As Chapters 6–8 indicate, income is the most important, and most justifiable, variable in IMF quota formulas and the UN scale. In 1982 the Fund adopted GDP as its income measure for quotas, while, beginning with the scale for 1998–2000, the UN instituted GNP as its measure for the scale of assessments. In 2000 the Quota Formula Review Group (QFRG) – a group of eight external experts charged with assessing quota formulas – reported its findings to the Fund. On the matter of the income measure, the QFRG confirmed that 'the single most relevant variable for measuring a country's ability to contribute to the IMF resources is GDP' (International Monetary Fund 2000a: 57). At about the same time, 'the [UN] Committee [on Contributions] concluded that GNP remained the least unsatisfactory income measure for calculating assessment rates' (United Nations 2003: 39).

Only the UN adjusts income for the debt burden of the country. In 1969 the Committee on Contributions initiated ad hoc downward adjustments in assessments for reason of external-debt servicing and amortisation. Debt adjustment to income became formalized with the 1986–1988 scale, with NNP replaced by debt-adjusted NNP. Interest payments on the debt are excluded from the adjustment, as they are not included in national product (NNP or, later, GNP). For a history and details of debt adjustment, see United Nations (2003: 40–41, 47). For criticisms of the adjustment, see Chapter 7, note 11.

External debt does not enter the Fund quota formulas. It was considered by the QFRG in the context not of income but of vulnerability to external economic disturbances, and rejected by that body (International Monetary Fund 2000a: 62). The Fund staff was more favorably inclined to inclusion of external debt in the formulas, but – as was the QFRG – was concerned with data adequacy (International Monetary Fund 2000d: 16).

Both the Fund and the UN use market exchange rates to convert income from domestic currency to a common currency, dollars. As argued in Chapters 6 (Section III.C and notes 18–20), 7 (Section III and note 9), and 8 (Section 4), an 'exchange-rate conversion bias' thereby acts to bias downward both the IMF quotas and the UN assessments of low per capita income countries, which are generally developing countries. Rationally, representatives of developing countries tend to criticize the first outcome and welcome the second – though both are due to the same factor. As Chapters 6–8 further argue, the use of purchasing power parity (PPP) in place of the exchange rate as the conversion factor would correct the bias. While, occasionally but rarely, the UN uses price-adjusted rates of exchange (PARE) in place of exchange rates, PARE is a relative-PPP concept – and it is absolute PPP that is required to correct the bias. (For a discussion of absolute and relative PPP from a methodological and historical standpoint, see Officer 2006.)

Exchange-rate conversion, that, for low per capita income countries, acts to reduce both IMF quota (presumed an unfavorable outcome for developing countries) and UN assessment (presumed a favorable outcome for developing countries) is expected on political grounds. In the IMF, there is weighted voting, largely according to level of quota (itself well correlated with economic size); in the UN, the rule is 'one country, one vote'). The section 'Net financing advantage,' below, examines the presumption that a higher IMF quota and lower UN assessment are unambiguously desirable for a member country of both organizations.

However, there are arguments in favor of the exchange rate and against PPP as the conversion factor. Regarding the UN scale, Broadbent observes that assessments are payable in U.S. dollars and that market exchange rates reflect the domestic-resource cost of acquiring dollars (foreign exchange). 'They [poorer countries] do not have the option of paying their UN contributions by buying US dollars at a PPP exchange rate' (Broadbent 1996: 83). From a data standpoint, according to Broadbent, the lag in PPP availability prevents their use in the UN scale.

The QFRG (2000a: 57–58) agrees with Broadbent. PPP is not an appropriate indicator of the ability to contribute to the IMF. Rather, a country's capacity to do so is determined by its ability to acquire foreign exchange at market exchange rates. Also, there are data problems with PPP. The QFRG sees the volatility of market exchange rates as a defect of their use, but suggests that this limitation could be overcome by averaging exchange-rate-converted income over a number of years, say, three years. Interestingly, such averaging has long been a feature of the UN scale. A two-year average began with the 1950–1951 scale, a three-year average with the 1954 scale, a seven-year average with the 1978–1979 scale, and a ten-year average with the 1983–1985 scale. Beginning with the 1995–1997 scale, an average of two sets of yearly averages has been taken (with the scale for 1998–2000 an exception), and one of the sets is at least a six-year average (see United Nations 2003: 46–8).

The Fund staff, also, argues that PPP-converted GDP would be a misleading indicator both of a country's ability to finance the Fund and of the potential need for drawings (International Monetary Fund 2000d: 13). Further, it discerns three types of PPP data limitations: lack of availability for all member countries, tendency to be out of date, and measurement error. The staff agrees with the QFRG that GDP should be averaged over several years, to iron out misleading GDP figures associated with exchange-rate volatility and the business cycle.

Jha and Saggar (2000) support PPP-converted GDP for IMF quotas, as do Kelkar, Yadav, and Chaudhry (2004). However, Rapkin and Strand (2005) see both merit and limitation in each method of conversion. Further, they state: 'the fact that a switch to PPP-based measurement of GDP would shift considerable IMF quota and votes from developed to developing countries is not in itself a sufficient reason to either oppose or favor such a change' (Rapkin and Strand 2005: 2007–8). Notwithstanding all these arguments, the present author continues to believe that a PPP-adjusted income for both the UN and IMF would enhance the logic and transparency of the process.

The UN alone has a low per capita income adjustment, which is an explicit progressive component in the scale viewed as taxation of member countries. For description and history of this element, see Chapters 7 (Section IV) and 8 (Section 4), and United Nations (2003: 41, 44–8). This adjustment shifts a substantial number of points (each point being one-hundredth of a percent) in the scale, as shown in Chapter 8 (Table 8.2) and Broadbent (1996: 84).

The IMF has never had such a variable in quota formulas. The QFRG (International Monetary Fund 2000a: 61) considers, and rejects, including either per capita income (with a negative sign) or population (with a positive sign) in its recommended quota formula. The unstated observation is that such variables would increase the quota share of developing countries at the expense of developed countries. The stated rationale for per capita income is that it proxies (inversely) the opportunity cost of holding international reserves, as it reflects the shortage of capital, and therefore enhances the demand for IMF quota (a contingent substitute for reserves). The justification for population is that it considers individuals and not just countries. However, as the Fund staff comments, neither variable is highly correlated with the ability to provide resources to the Fund or with the potential need for balance-of-payments financing (International Monetary Fund 2000d: 16).

On its part, the IMF incorporates several variables in quota formulas that, justifiably, play no role in the UN scale of assessments. Openness – proxied by current-account receipts and payments (see Chapter 6, Section II.B), is a measure of economic size that can indicate both the ability to contribute resources to the Fund and the potential need for drawings. The QFRG asserts that openness is best measured by value-added in the tradables section, but such data are not available (International Monetary Fund 2000a: 58–59).

An important way in which openness variables could be improved is by incorporating the financial (capital) account of the balance of payments as well as the current account. The Fund has long considered inclusion of capital-account variables in the quota formula – see International Monetary Fund (2000b: 72–5). For discussion of alternative openness variables, based on international investment position (perhaps proxied by international investment-income flows), see International Monetary Fund (2000b: 6–9).

The ratio of current-account receipts to GDP is a multiplicative term in some quota formulas (see Chapter 5, Section II.B). Though on the surface a reasonable indicator of openness, this ratio can give rise to the anomalous result that an increase in GDP involves a reduced calculated quota. The anomaly is worsened by the facts that (1) receipts are measured in gross terms, while GDP is a net (value-added) concept; and (2) the ratio, even if measured properly, is an imperfect indicator of openness. Although the multiplicative nature of the ratio makes the quota formula unnecessarily complex, it is the ratio – not its nonlinear role – that gives rise to the anomaly. On all this, one may consult the International Monetary Fund (2000a: 61–4; 2001b: 12).

Variability of current-account (formerly commodity-export) receipts has been a component of every operational quota formula since Bretton Woods. As the QFRG states: ‘The single most relevant variable for measuring a country's vulnerability to external economic disturbances is the variability of its international receipts' (International Monetary Fund 2000a: 59). This variable can also incorporate the ability of a country to contribute resources to the Fund. The measure could be improved by including autonomous net inflow of capital – a QFRG recommendation supported within the Fund (see International Monetary Fund 2000d: 14; 2002: 10). However, the Fund staff notes severe data problems in computing capital-account variables (International Monetary Fund 2000d: 17–19). Inclusion of the capital account in the variability measure is also discussed in International Monetary Fund (2001b: 15–20).

A history of the variability measure and an outline of alternative measures of variability (including the addition of net long-term capital flows to current receipts) is provided in International Monetary Fund (2000b: 95–103). Jha and Saggar (2000: 589) argue that the variability measure is defective in not taking account of the mean. Presumably, they would prefer the ratio of the standard deviation to the mean as the measure rather than simply the standard deviation (from trend). However, the Fund's computation makes better sense, because it allows for a country-economic-size effect.

A final variable in quota formulas that measures economic size and pertinent only to the Fund is international reserves. This variable is an indicator of the capacity to provide resources to the Fund. Justification of this variable is in Chapter 6 (Section III.C); for discussion of limitations of the variable, see International Monetary Fund (2000d: 15; 2002: 10).

Review of literature on country-group financing

Chapters 6–8 provide quantitative examination of IMF quotas and UN assessments by country group, with special attention to developing countries. There exists a small, but pertinent, literature on the topic; it is outlined here in the context of Chapters 6–8.

It is important to distinguish between *actual* and *normative* quotas or assessments as the subject of empirical work. Consider *actual* financing shares first. Two approaches are discernible. The simpler technique is to show country-group distribution of financing, with no attention paid to determination of the distribution. This approach is seen in Chapter 7 (Section II), where the distribution of UN assessments by country group is shown for the three scales spanning 1986–1994. De Vries (1985: 536–8) and Boughton (2001: 854–56) do the same for IMF quotas, together incorporating eight dates over 1945–1989. The weakness of this, first, approach is that it is purely descriptive.

The alternative technique involves explaining actual quotas via cross-section (international comparison) regressions. The unit of observation is the individual country, but what is of interest is the implication for country groups. In Chapter 6 (Section VI.B), quotas are determined by regression on real (constant-price, PPP-converted) GDP, trade (receipts plus payments on goods-and-services account), official reserves, population, and a constant term modified by dichotomous variables for 'industrial countries' and 'other developed countries.' A moving cross-section regression is performed for eight dates over 1950–1985. The main finding is that there is increasing favoritism toward developing countries over time.

The UNCTAD Secretariat (1987: 533–7) uses this technique for one year (1979), regressing quota on income, imports, volatility of exports, and a constant term. The regression is run for different groups of countries: all members, later (joined after 1950) members, founding members, founding members excluding four large-quota countries (China, France, United Kingdom, United States). The QFRG (2000a: 43–53; 2000c: 55–70) estimates a large number of equations with dependent variable quota of the Eleventh General Review. Independent variables include GDP (converted alternatively via exchange rate and PPP – the former giving rise to five-year-average GDP), reserves with gold valued at market (rather than official) prices, current-account receipts and payments, and various measures of variability. Relative contributions of variables to quotas, and estimated quota shares, are shown by country group.

Chapter 8 (Section 7) examines the UN actual assessment, using the regression technique. The assessment/income ratio is the dependent variable. Explanatory variables are per capita real income and dichotomous country-group variables. Cross-section regressions are run for all meaningful scale periods 1947–1994. Excluded from the samples are the ceiling country (United States), floor countries, and China (until 1980–1982).

Turning to empirical work on *normative* financing, the simpler approach is to compare the actual distribution of quota or assessment (by country group) with the corresponding normative distribution, where the normative outcome is obtained via some formula or algorithm advocated by the researcher. Chapter 6 (Section 5) is an example of this procedure, with the developing-country percentage of actual quotas compared with the developing-country percentage of PPP-converted real GDP, income, trade, and reserves. Computations are made for eight dates over 1950–1985. International Monetary Fund (2001b: 11) shows shares of quotas by country group with corresponding shares of the variables entering quota formulas: GDP, current-account receipts, current-account payments, variability of receipts, and reserves.

Kelkar, Yadav, and Chaudhry (2004) advocate quota shares strictly proportional to PPP-converted GDP. They compare then-existing country group quota shares with shares of GDP (1997–1999 average). Of course, developing countries as a group gain – and advanced economies lose – quota share.

The QFRG (International Monetary Fund 2000a: 5–6, 64–5) recommends a new quota formula: quota is the weighted sum of three-year-averaged, exchange-rate-converted GDP and a measure of external variability (preferably of current-account receipts plus net long-term capital flows). The suggested weight is 2/3 for GDP and 1/3 for variability. Interestingly, the QFRG did not itself compute quotas or quota shares emanating from its formula. This task was undertaken by the Fund itself (International Monetary Fund 2000d, 2001a).

Using data for the Eleventh General Review and excluding long-term capital flow from the variability measure, due to lack of data for all countries, country-group results are presented for actual quota shares, QFRG-formula quota shares, and existing five-formula quota shares. Incorporating capital flow data, available for a subsample of countries, country-group results are exhibited for current quota shares, calculated Eleventh Review shares, calculated shares from traditional five formulas, and calculated shares from the QFRG formula (the latter with original weights and reversed weights). Results of the two studies are consistent. Principally, 'the [QFRG] proposed formula could result in a significant shift in quota shares toward the advanced, large, and relative closed economies' (International Monetary Fund 2000d: 9).

Important for this result is the high weight of GDP in the QFRG formula and the exclusion from the formula of any measure of openness. It is reasonable to hypothesize that replacement of exchange-rate-converted GDP by PPP-converted GDP would have acted as a correction to this shift.

Excluded from this survey are Fund computations of quotas that would result from alternative Fund formulas or alternative variables in existing Fund formulas (for example, International Monetary Fund 2001b, 2002).

Jha and Saggar (2000) use PPP-based GDP and capital flow as variables in quota formulas, among other changes and alternatives. They exhibit results only for India, but report quota gains for developing countries generally (with the exception of oil-exporting countries and some small countries). UNCTAD Secretariat (1987) examines the effect of alternative variables and formulas on developing-country quota shares, but makes no normative statement. Rapkin and Strand (2005) compare actual quotas with calculated quotas from a variety of sources: Eleventh General Review, QFRG formula, and two formulas advocated by Vijay Kelkar (employing PPP-GDP, with population also in one formula). Tables are shown for the following country categories: high-income, OPEC, upper-middle-income, lower-middle-income, and low-income. No summary country-group table is provided.

Considering the UN, Chapter 7 offers a country-group comparison of the actual scale with a normative scale based on (i) constant-price PPP-converted GDP and (ii) the UN low-per-capita-income formula applied to per-capita GDP adjusted for development level. This normative scale has the advantage of simplicity, transparency, and overt favoritism to developing countries. The analysis is done for the three scales over 1986–1994. A related, but somewhat different, investigation is performed in Chapter 8 for all meaningful scales over 1957–1994.

General Accounting Office (1994) presents percentage of exchange-rate-converted GNP as the normative assessment scale. The computation is made alternatively including and excluding a 25-percent ceiling. Country groups distinguished are the four peacekeeping groups: permanent members of the Security Council, developed countries not such permanent members, least developed countries, and other developing countries (see Chapter 7, note 1, and Chapter 8, note 2). Results are shown for both the UN regular scale and the peacekeeping scale.

Klein and Marwah (1997) investigate several nonlinear assessment formulas, taking the form of per-capita assessment a function of per capita income. The functions are restricted to pass through a ceiling, a floor, and sometimes also median per capita income. GDP conversion is performed alternatively via exchange rate or PPP. Results are shown by country group, but only for some cases.

Net financing advantage

Introduction

The conflicts generated by the allocation of UN and IMF funding to member countries are intense. Financing is important, because it permits these organizations to operate. However, financing in terms of shares is a zero-sum game, which is the source of country conflict. The large number of countries in the organizations leads naturally to a coalescence of members into groups (especially developing versus developed countries) for purposes of negotiating the country allocation of financing. This section examines the first half-century of experience in UN/IMF financing from that perspective.

The distinction is made between 'gainers' and 'losers' in the country distribution of international-organization financing. What makes this distinction possible is a model of 'neutral bargaining outcome,' developed here to provide a norm. What makes the norm possible is joint consideration of country-distribution outcomes of UN and IMF financing and the fact that 'financing' of the UN and IMF has different connotations. UN financing is primarily private cost of public good (maintenance of peace and security, international approaches to other issues); IMF financing covers predominantly a set of private benefits. (This is not to deny that the UN also provides some private goods, such as development assistance to individual members, and that the Fund produces some public goods, such as multilateral surveillance of exchange rates.)

Measure of country net advantage

Let A_i denote the UN assessment rate of country i (where $\Sigma A_i = 100$), S_i the country's IMF quota in absolute (dollar or SDR) terms, and Q_i its percentage share in total Fund quotas (where $Q_i = 100 \cdot (S_i/\Sigma S_i)$ and $\Sigma Q_i = 100$). Given the total of Fund quotas, ΣS_i, a higher percentage quota corresponds to a higher absolute quota ($dQ_i/dS_i > 0$). Note that a fixed total of Fund quotas is not a mere arithmetic convenience. Total quotas, sometimes called the 'size of the Fund,' required determination at Bretton Woods and are re-set at the reviews of Fund quotas that occur periodically.

Let U_i represent the utility of country i. Then the attributes of the assessment rate and percentage quota are used to demonstrate that $dU_i/dQ_i > 0$ and $dU_i/dA_i < 0$. It should be noted that just as IMF quota share corresponds to UN assessment rate, so total quotas correspond to the UN budget. It is fair to say that the size of the Fund in the form of total quotas and the size of the UN in the form of its regular and peacekeeping budgets are as controversial issues as are the country percentage distribution of quotas and the UN assessment scale. However, it is only the latter that are the concern here (and of Chapters 6–8).

Transfer of resources

Because the UN produces public goods, it obtains revenue by taxing member countries annually via the scale of assessments, with amount payable equal to one hundredth the product of the country's assessment rate and the UN regular budget. (As observed above, the peacekeeping budget is financed via a derivative assessment scale.) Generally, assessments are due in dollars: a budgetary cost for the United States, a foreign-exchange cost for other countries. The free-rider problem plagues UN financing: for a given level of UN spending, a country's benefits are independent of its assessment. Therefore, from both the cost and benefit perspectives of resource transfer, $dU_i/dA_i < 0$.

In contrast, the IMF does not tax members for its administrative and operational expenses, relying principally on charges related to member use of the Fund's resources. Analogous to the UN assessment is the member's subscription (or subscription increase), equal to its quota (or quota increase) in amount and payable to the Fund partly in reserve assets (originally gold, now SDRs or widely accepted currencies) and partly in domestic currency – the reserve-asset proportion generally 25 percent for quota increases, though varying for initial quotas.

The apparent burden of a quota subscription is the reserve-asset component, because the member relinquishes reserves to the Fund, which at first glance makes $dU_i/dQ_i < 0$. However, as an offset, the country's reserve position in the Fund increases in amount equal to its reserve-asset subscription, and this position since 1952 in practice (and since 1969 in law) constitutes an automatic right to make drawings from the Fund. Therefore subscribing to its quota (or quota increase) changes not the total amount of the member country's official reserves, only the composition of these reserves. The only cost is a reserve-asset composition different from that which would be otherwise desired by the country. In 1978, even the opportunity cost of interest forgone was eliminated by the Second Amendment to the Fund's Articles of Agreement. 'Hence there was no longer any ex ante cost of investing resources in the Fund' (Boughton 2001: 855). Indeed, Boughton uses the term 'imagined costs' to describe ex ante cost. And any ex post cost could just as well be a benefit (Boughton 2001: 855, n. 15, and 900–4).

The domestic currency portion of the subscription need not involve even a budgetary cost, at least immediately. This portion is typically held by the IMF in depository accounts at members' central banks (International Monetary Fund 2000a: 10). However, the Fund permits it to be subscribed in non-negotiable, non-interest-bearing notes payable on demand. Subsequent Fund use of the currency, while entailing a budgetary cost, correspondingly enhances the member's reserve position (automatic drawing right) in the Fund and thus its official reserves.

On balance, while it cannot be denied that subscribing to a quota in itself involves dU_i/dQ_i with negative rather than positive sign, it is fair to say that the magnitude of dU_i/dQ_i (by reason of subscription) is much smaller than a cursory judgment would suggest.

Voting power

Voting rights in the IMF are given by the formula $V_i = 250 + S_i/100{,}000$, where V_i is the number of votes of country *i*. The 250 'basic votes' aggregated over all member countries have always constituted a small proportion of total votes, and have declined over time (see 'General characteristics of financing the IMF and UN,' above). Weighted voting applies everywhere in the Fund, irrespective of the Fund body in which the decision is made. Clearly, the voting formula in itself makes $dU_i/dQ_i > 0$.

In the UN, voting power is independent of assessment, with each member having one vote in the General Assembly and the five permanent members of the Security Council singly having veto power on nonprocedural matters in the Council. This voting structure in itself yields $dU_i/dA_i = 0$.

Share of output

A member's quota determines the limits to its drawings (borrowing) of other members' currencies and SDRs from the Fund, with maximum access to the Fund's financial resources, under various facilities, expressed in terms of percent of quota. (There are exceptions – see 'General characteristics of financing the IMF and UN,' above). Also, SDRs, when created by the Fund, are allocated to members strictly as a percentage of quota (and are not 'drawn,' but usable largely at a country's discretion). In contrast, UN output is composed largely of public goods, determined independently of assessment. So this element, also, makes $dU_i/dQ_i > 0$ but $dU_i/dA_i = 0$.

Prestige

It is natural for a country to view its national prestige as affected positively by its share of IMF quotas. This phenomenon was present at the Bretton Woods Conference (Mikesell 1994: 35) and no doubt has continued throughout the Fund's history. Before the first scale of assessments was determined, in a resolution of 13 February 1946, the General Assembly declared that 'some members may desire unduly to minimize their contributions, whereas others may desire to increase them unduly for reasons of prestige.' This predicted symmetry did not occur: 'The fear that Member States may wish to increase their contributions unduly has proved unfounded' (United Nations Secretariat 1989: 5). Once again the same result: prestige, too, involves $dU_i/dQ_i > 0$ and $dU_i/dA_i = 0$.

The four financing attributes together imply ambiguously that $dU_i/dA_i < 0$. In contrast, the sign of dU_i/dQ_i is uncertain from a mathematical standpoint, because three forces manifest $dU_i/dQ_i > 0$ and one $dU_i/dQ_i < 0$. However, it is argued above ('Transfer of resources') that the force yielding $dU_i/dQ_i < 0$ is weak and uncertain.

More important, revealed preference of IMF members provides strong empirical evidence that almost always $dU_i/dQ_i > 0$. On this revealed preference, Chapter 6 (note 4) should be consulted, complemented by the following facts. Bargaining at general reviews is intense, with countries concerned with their relative quota shares as well as the total quota increase and its relationship to their shares – well demonstrated in the Fund's 'inside histories' (see 'Data and information sources,' below). From the 1958–1959 quota review (the first review resulting in an increase in quotas) through the Ninth General Review (the last review incorporated in this chapter), there were only six cases of countries electing not to participate at quota reviews (and hence becoming ineligible for quota increases) and only in 27 instances did participating countries consent to none or only part of their proposed quota increase – dwarfed by a total of over 800 observations. Eleven of the 33 contrary observations (involving China, Cambodia, Cuba, and Iran) were due to political considerations, and nine of the remainder (all under the Ninth General Review) concerned countries forbidden to consent to quota increases because in arrears to the Fund. However, four of the nine countries settled their arrears and consented at a later date – see Boughton (2001: 873, n. 61). The QFRG counts only 22 meaningful instances in which a member did not consult to its proposed maximum quota increase, from the 1958–1959 review through the Ninth General Review (International Monetary Fund 2000a: 30).

Therefore it is reasonable to assume $dU_i/dQ_i > 0$ and $dU_i/dA_i < 0$. Assume that country i attempts to maximize U_i. It does this by maximizing Q_i and minimizing A_i. Define D_i^u, the 'unadjusted financing differential' measure as $D_i^u = Q_i - A_i$. Though every country i seeks to maximize D_i^u, this is a zero-sum game. Assuming universal membership in the UN and IMF, so that (Q_i, A_i) exist together for any i, then $\Sigma D_i^u = 0$.

Therefore the norm value of D_i^u is zero. Other things being equal, any country i expects $D_i^u = 0$. If $D_i^u > 0$ ($D_i^u < 0$), the country has a net advantage (disadvantage) *relative to other countries* in the financing of the UN and IMF, and it is assumed that the gain or loss increases with the magnitude of D_i^u, that is, $dU_i/dD_i^u > 0$. The underlying assumption, that an added percentage-point of IMF financing has the same utility as a reduced percentage-point of UN financing (always relative to other countries, rather than in an absolute sense) simplifies the model, by restricting the game to be zero-sum. One could allow U_i to vary with the levels of Q_i and A_i rather than just with their difference; but this would unduly complicate the analysis.

Samples over time

The UN Charter specifies that the General Assembly (GA) has the responsibility of approving the UN budget and determining the apportionment of UN expenses among its members, and the GA does so by majority vote. The

Assembly appoints a Committee on Contributions (CC) to make recommendations regarding the scale of assessments, and provides constraining directives for the CC's work. The CC actually reports to the GA's Administrative and Budgetary Committee (Fifth Committee, FC), where, with each member having one vote, draft resolutions are recommended to the GA. Because their membership is identical and universal, the outcome in the FC is a preview of that in the GA, where the vote occurs only days later.

Major decisions on the scale of assessments occur when all member countries are considered together and the entire scale is determined in a general-equilibrium setting (wherefore generally $\Sigma A_i = 100$). These decision periods encompass the meetings of the CC, deliberations of the FC on the CC recommendations, and action of the GA. Each decision period results in a scale period (the year or years for which the new scale is applicable) or scale periods (if different scales are determined for future years). There were 25 such decision periods from the inception of the UN to 1994, resulting in scale periods from 1946 through 1997. These episodes are to be conjoined with corresponding decision periods of IMF quota determination.

Originally, an 80 percent majority of the voting power in the Board of Governors (BG, ultimate authority in the Fund) was required for any change in quotas. In 1969 this was changed to 85 percent for a general review of quotas, extended in 1978 to any change in quotas. In practice, all changes in quotas are considered in the Executive Board (EB, the body responsible for conducting the Fund's operations). For general reviews of quotas, the process begins with consideration in a Committee of the Whole of the EB, with formal recommendations made by the EB to the BG.

Again, the pertinent quota decisions involve joint consideration of members' quotas. The quotas of countries that participated in the Bretton Woods Conference and that joined the Fund by the end of 1946 were as negotiated at the Conference. Originally a quinquennial review of the quotas of all members was mandated; in 1969 this was changed to a 'general review' at intervals of not more than five years. Additional reviews, at other than the stipulated times, are always permitted. From 1946 (the beginning of the Fund's operation) to 1995, there were eleven completed reviews, one of which (1958–1959) was non-mandated. Along with the Bretton Woods Conference, these are the episodes at which the entire quota structure – both the Fund total and the country distribution of quotas – is considered and, if so decided, altered. The corresponding decision periods (after 1944, for the Bretton Woods Conference) are the intervals from the first meeting of the EB Committee of the Whole to the BG final resolutions on the quota review. Therefore a set of pairings (UN, IMF) is to be selected from 25 UN and 12 IMF sample points.

The strategy is to synchronize dates at which a new quota distribution was established for IMF membership and a new assessment scale set for UN membership. A sample pair involves decisions of both bodies at as close as possible to the same time.

What is of interest are the organizations' decisions, not the subsequent behavior of countries. Thus the relevant datum for a UN member is its assessment rate in the GA resolution rather than the portion of this rate that it in fact pays. Similarly, the recorded quota for an IMF member (the maximum amount that the country can obtain) is that proposed in the BG resolution rather than the (rarely, but possibly, lesser) amount to which the country consents.

Nine samples are obtained for the first half-century of the joint existence of the two organizations. They are identified by the corresponding IMF and UN decision dates in the first two columns of Table 9.1 and, alternatively, by the joint time span of these decision dates in the first column of Table 9.2. The 1944–1946 sample encompasses the quotas determined at the Bretton Woods Conference in July 1944 and the initial UN scale of assessments, adopted by the GA in 1946. The first, second, and third quinquennial review of quotas (in 1949–1951, 1954–1956, and 1958–1960) involved no change in quotas, thereby excluding samples based on them.

Table 9.1 Characteristics of samples

Decision date		*Common membership (percent)*	
IMF	*UN*	*IMF*[a]	*UN*[b]
1944	1946	99.99	94.95
1958–60	1961	94.52	77.26
1964–66	1964–65	94.02	78.28
1969–70	1970	93.98	79.80
1974–76	1976	99.35	82.35
1977–78	1979	99.56	82.91
1982–83	1982	99.49	84.60
1987–90	1991	99.99	88.21
1992–95	1994	98.29	99.74

a Sum of quota percentages of countries that are also UN members.
b Sum of rates of assessment of countries that are also IMF members

Table 9.2 Adjusted financial differential, by country group (percentage points)

Sample	*United States*	*Other developed*	*Eastern Europe*	*Developing*
1944–46	–10.76	3.63	7.78	–0.65
1958–61	–12.00	3.85	—	8.15
1964–66	–13.70	4.02	—	9.68
1969–70	–13.97	2.82	—	11.15
1974–76	–8.63	–7.30	0.31	15.62
1977–79	–8.51	–11.77	0.37	19.91
1982–83	–9.50	–12.79	0.68	21.61
1987–91	–8.72	–12.53	0.92	20.33
1992–95	–6.33	–13.81	–0.12	20.26

However, the non-mandated review of 1958–1959 in effect replaced the third quinqennnial review. The four BG resolutions that ended the review, together with new quotas for four countries approved in 1960, constitute the IMF component of the 1958–1961 sample. The UN component is the scale for 1962–1964, adopted by the GA in 1961. The 1964–1966 sample is based on (1) the two BG resolutions of the fourth quinqennial review (plus allowing quota increases under the 'compensatory financing facility' to serve as the base quota for seven countries that applied for increases in 1965 after the EB report to the BG) and (2) the scale of assessments for 1965, approved by the GA in that year.

The next five samples, 1969–1970 to 1987–1991, incorporate (1) on the IMF side, the BG resolutions concluding the Fifth through Ninth General Reviews, respectively, and (2) on the UN side, the scale of assessments approved by the GA for 1971–1973, 1977, 1980–1982, 1983–1985, and 1992–1994, respectively. The final sample, 1992–1995, involves (1) quotas at the end of 1994, and (2) the scale of assessments for 1997, approved in 1994. The Tenth General Review, terminated by BG resolution in January 1995, left quotas unchanged. However, in 1992 the 15 states that devolved from the breakup of the USSR joined the IMF. This produced Eastern Europe as an important country group in the Fund, and therefore in the sample, for the first time since 1944–1946. Fitting Eastern European quotas into the existing quota structure had at least the partial impact of a general review.

Country conflicts and resulting groupings

To establish country groupings, it is useful to explore the history of country conflicts in the organizations. For the UN, the discussion in Chapter 8 (Section 6) may be supplemented by a fresh examination here. Looking at the work of the CC and debate in the FC, three country group conflicts – and four country groups – are apparent in UN history. First, the United States, opposed by Eastern Europe (during the Soviet-bloc era) and by developing countries, has always sought to reduce its, the largest, assessment. Second, in the 1950s and 1960s, Eastern Europe protested vigorously increases in its assessments, which were defended principally by the United States and other developed countries.

Third, beginning in the early 1950s but especially from the late 1960s, developing countries complained bitterly whenever assessments of some developed countries were reduced while assessments of some developing countries were increased. Naturally, they were opposed by developed countries. By force of numbers, the developing country group acquired and strengthened a majority position in the GA. From 1951 onward, many a GA resolution instructed the CC to develop the scale of assessment giving consideration to the developing countries or to low-per-capita-income countries. Debate in the FC reveals that the developed-versus-developing country-group conflict has been the dominant financing issue in the UN since the second half of the 1960s (see, for example, United Nations Secretariat 1989: 3).

In the IMF, Eastern Europe was not a player in the conflicts at quota reviews (at least in the first half-century of the Fund). Also, the weighted-voting structure geared to economic size weakens the power of the developing countries. What is most apparent, beginning with the first quinquennial review, was the attempt by the United States to maintain its, again largest, share of total IMF quotas and to minimize slippage over time. Conflict with potentially higher-quota members, particularly other developed countries, is evident.

Recorded conflict between the developed and developing countries over the distribution of quotas began in the late 1960s, with developing countries pressing for a larger percentage share and developed countries resisting any reduction in their share (de Vries 1976, vol. 1: 303–305; 1985, vol. 1: 511–512).

In sum, four groups – the United States, other developed countries, Eastern Europe, and developing countries – encompass the principal country-group conflicts in financing the UN and IMF, for the first half-century of these organizations. For country allocation to the groups, guidance is found in the United Nations *Statistical Yearbook*. 'Other developed countries' embody OECD membership (excluding the United States and Turkey) plus South Africa plus, for the 1974–1995 samples, Israel and Malta. Eastern Europe is the traditional Soviet bloc or successor countries, with Yugoslavia in the 1944–1946 sample.

Financing-differential results

The unadjusted-financing-differential measure, $D_i^u = Q_i - \mathrm{A}_i$, is defective even when restricting a sample to members of both the UN and IMF, because, with membership non-universal, $\Sigma Q_i \neq \Sigma A_i$, as shown in the third and fourth columns of Table 9.1, wherefore $\Sigma D_i \neq 0$. For equal weighting of Q_i and A_i, the correct measure is the 'adjusted financial differential,' $D_i = Q_i - k{\cdot}A_i$, where $k = \Sigma Q_i/\Sigma A_i$ is specific to the sample at hand, which yields $\Sigma D_i = 0$.

Table 9.2 presents the distribution of D_i summed by country group. The United States is always a net 'subsidizer' of other groups in joint financing of the UN and IMF. However, U.S. 'loss' drops precipitously in 1974–1976 and again in 1992–1995. Other developed countries receive a 'subsidy' until 1974–1976, when in two jumps their 'loss' exceeds that of the United States. Eastern Europe is a big 'gainer' in 1944–1946, but almost even thereafter. After 1944–1946, developing countries are the heavily subsidized group, with the net gain steadily increasing to about 20 percentage points in 1977–1995.

Data and information sources

All known information about the scale of assessments is self-reported by the UN (see United Nations Secretariat (1989) and United Nations, *Report of the Committee on Contributions*, annual; *Scale of Assessments for the*

Apportionment of the Expenses of the United Nations: Report of the Fifth Committee, annual plus occasional addenda; and *Yearbook of the United Nations*, annual).

Almost all information about quotas comes from the IMF itself, especially its *Annual Report*, three 'inside-histories' (Horsefield et al. 1969; de Vries 1976 and 1985, Boughton 2001), and many miscellaneous publications: Altman (1956), Gold (1974), Edo (1978), Chandavarkar (1984), and International Monetary Fund (2000a, 2000b, 2000d, 2001b, 2001c, 2003, 2004, 2005). Independent outside information pertains exclusively to quotas determined at the Bretton Woods Conference, where the IMF was negotiated (see Mikesell (1994) and the references in Altman (1956: 141–2, n. 26)).

Data sources for A_i for the nine samples are *Yearbook of the United Nations*, 1946–1947: 219; 1961: 573–4; 1965: 681–3; 1970: 860–3; 1976: 898–900; 1979: 1207–8; 1982: 1418–19; 1991: 879–80; United Nations (1994: 18–21).

For Q_i, the sources are: (1): Horsefield et al. (1969, vol. 3: 210). (2): International Monetary Fund, *Summary Proceedings*, 1960: 158–61, 169; *Annual Report* 1959: 188–9; Horsefield et al. (1969, vol. 2: 378–80; vol. 3: 429–30). (3): International Monetary Fund, *Summary Proceedings* 1965: 252–7; 1966: 236–8; Horsefield et al. (1969, vol. 3: 463, 465). (4)–(6): de Vries (1976, vol. 2: 271–2; 1985, vol. 3: 250–1, 262–3). (7)–(8): International Monetary Fund, *Annual Report* (1983: 140–2; 1990: 104). (9): International Monetary Fund, *International Financial Statistics*, February 1995: 10–12.

References

Altman, O. L. (1956) 'Quotas in the International Monetary Fund', International Monetary Fund *Staff Papers*, **5**: 129–50.

Boughton, J. M. (2001) *Silent Revolution: The International Monetary Fund 1979–1989*. Washington: International Monetary Fund.

Broadbent, S. (1996) 'Financing the United Nations: International Taxation Based on Capacity to Pay', *National Institute Economic Review*, **157**: 77–89.

Chandavarkar, A. C. (1984) *The International Monetary Fund: Its Financial Organization and Activities*. Washington: International Monetary Fund.

de Vries, M. G. (1976) *The International Monetary Fund 1966–1971: The System Under Stress*, 2 vols., Washington: International Monetary Fund.

—— (1985) *The International Monetary Fund 1972–1978: Cooperation on Trial*, 3 vols., Washington: International Monetary Fund.

Edo, M. E. (1978) 'Multiformula Method Adds Flexibility in the Calculation of Members' Quotas', IMF *Survey*, June 5: 166–68.

General Accounting Office (1994) *United Nations: How Assessed Contributions for Peacekeeping Operations Are Calculated*, Washington.

Gold, J. (1974) *Membership and Nonmembership in the International Monetary Fund*, Washington: International Monetary Fund.

Horsefield, J. K. and others (1969) *The International Monetary Fund 1945–1965: Twenty Years of International Monetary Cooperation*, 3 vols., Washington: International Monetary Fund.

International Monetary Fund (2000a) *Report to the IMF Executive Board of the Quota Formula Review Group*, available on www.imf.org.
—— (2000b) *External Review of the Quota Formulas – Annex*, available on www.imf.org.
—— (2000c) *Report to the IMF Executive Board of the Quota Formula Review Group: Statistical Appendix – Part A*, available on www.imf.org.
—— (2000d) *Staff Commentary on the External Review of the Quota Formula*, available on www.imf.org.
—— (2001a) *External Review of Quota Formulas: Quantification*, available on www.imf.org.
—— (2001b) *Alternative Quota Formulas: Considerations*, available on www.imf.org.
—— (2001c) *Financial Organization and Operations of the IMF*, 6th ed., Washington.
—— (2002) *Alternative Quota Formulas – Further Considerations*, available on www.imf.org.
—— (2003) *Quota Distribution: Selected Issues*, available on www.imf.org.
—— (2004) *Quotas – Updated Calculations*, available on www.imf.org.
—— (2005) *Quotas and Voice – Further Considerations*, available on www.imf.org.
Jha, R. and M. K. Saggar (2000) 'Towards a More Rational IMF Quota Structure: Suggestions for the Creation of a New International Financial Architecture', *Development and Change*, **31**: 579–604.
Kelkar, V., V. Yadav and P. Chaudhry (2004) 'Reforming the Governance of the International Monetary Fund', *World Economy*, **7**: 727–43.
Klein, L. R. and K. Marwah (1997) 'Burden Sharing for Support of the United Nations', in J. Boulden (ed.) *Aurora Papers 30: Unconventional Means of UN Financing*. Ottawa: Canada.
Mikesell, R. F. (1994) *The Bretton Woods Debates: A Memoir*. Princeton: International Finance Section, Princeton University.
Officer, L. H. (2006) *Purchasing Power Parity from Ancient Times to World War II*, in R. Whaples (ed.) *EH.Net Encyclopedia*.
Rapkin, D. P. and J. R. Strand (2005) 'Developing Country Representation and Governance of the International Monetary Fund', *World Development* **33**: 1993–2011.
UNCTAD Secretariat (1987) 'Determination of Quotas and the Relative Position of Developing Countries in the International Monetary Fund', in S. Dell (director) *The International Monetary System and its Reform*, Amsterdam: North-Holland.
United Nations (1994) *Report of the Committee on Contributions*, 54th session. New York: United Nations.
—— (2003) *Report of the Committee on Contributions*, 63rd session. New York: United Nations.
United Nations Secretariat (1989) *Evolution of the Methodology for the Scale of Assessments and its Current Application*, New York: United Nations.

Part III

Monetary history

10 The bullionist controversy

A time-series analysis*

> The so-called 'bullionist' controversy . . . was probably the most important Foreign Exchange controversy for all time.
>
> (Einzig, 1970, p. 202)

Summary

The Bank Restriction Period (1797–1821) was the sole British experience with a paper standard and floating exchange rate to 1914. Contemporary observers disagreed vigorously about the relationships among the price level, exchange rate, and money supply (represented by Bank of England notes). On one side were the bullionists, comparable with modern monetarists. They argued that Bank of England notes determined the price level, which then determined the exchange rate. On the other side were the antibullionists, who emphasized non-monetary influences on the exchange rate, which helped to determine the price level, which was then accommodated by Bank of England note circulation. A modern bullionist theory partially bridges the gap between the two camps; its focus is on Bank of England note circulation determined by demand and supply. Superior data and more sophisticated techniques than those of previous authors are employed to test the three positions. Evidence is preponderantly in favour of the antibullionist approach, which ironically is out of fashion in modern macroeconomics.

Introduction

The Bank Restriction Period provides the earliest example of divergent monetarist/nonmonetarist approaches to macroeconomics under a paper standard and floating exchange rate. During the Bank Restriction Period, Britain's *domestic* gold standard was replaced by a paper standard, because all banks refrained from making cash payments, that is, redeeming their notes (and deposits) in gold coin. Britain's *international* gold standard was replaced by a floating exchange rate, because conventional gold points were inoperative: gold could not be obtained at a fixed price in domestic currency, as above, nor could gold bullion generally be sold for domestic currency at a meaningful official price. The gold price at the Mint (£3 17s. 10½d. per standard ounce), combined with waiting time, was too far below the market price for there to be private customers, except in 1817–1820, when favorable payment arrangements were temporarily in effect. Though the Bank of England's payment was speedier, its normal buying price (£3 17s. 6d.) was too low to be operative, and Bank of England purchases of gold took place only at higher prices, aligned with the market.[1]

The Bank Restriction Period began on 27 February 1797, with implementation of an Order in Council prohibiting the Bank of England from making cash payments, formalized by the Bank Restriction Act of 3 May 1797. Full resumption of specie payments, and therefore return to the gold standard, occurred on 1 May 1821.[2]

The Bank Restriction Period gave rise to 'the bullionist controversy', the most famous monetary debate in the history of economic thought. A measurement issue provided the nomenclature: the 'bullionists' asserted, and the 'antibullionists' denied, that the premium on gold bullion correctly gauged the depreciation of the paper pound. However, substantive topics – determination of the exchange rate and price level, and the behavior of the Bank of England – dominated both the contemporary debate and the interest of post-Bank Restriction Period writers. Here the bullionists adopted a clear monetarist approach and the antibullionists a decidedly nonmonetarist position.

The survey literature of the bullionist debate, as well as the writings of the contemporary protagonists, do not explicitly present the competing bullionist and antibullionist models as chains of causation. So it is not surprising that existing empirical testing considers the various hypotheses only separately and individually, resulting in purely bivariate testing of the models. Also, previous researchers have either lacked the advantage of good data or made inappropriate data selection.

In this Chapter the first task is to present the contemporary bullionist and antibullionist models, along with a modern bullionist alternative, as chains of causation linking individual hypotheses. Then time-series analysis is used to test the competing models in multivariate form, with careful attention to data collection and construction of variables. The empirical results provide strong, though not uniform, support for the antibullionist position.

Bullionist and antibullionist models

Methodology and notation

There is no need to provide another comprehensive survey of the bullionist debate.[3] The better procedure is to exposit each side by a testable model. While there were certainly nuances in the positions of contemporary authors, a general model for each side is readily discerned from the contemporary and survey literature. In Viner's (1937, pp. 120, 127) words, the 'essential doctrines of the bullionists' and of the antibullionists are delineated in contrasting models, abstracting from 'qualifications conceded by the bullionists' and antibullionists.

Consider the following notation:

BN Bank of England notes ('Banknotes') in circulation
BP Balance of payments (positive if deficit, negative if surplus)
ER Exchange rate (price of pound in terms of foreign currency), with inverse denoted as ER^{-1}
HR State of harvest (inversely related to quantity and quality)
ME External military expenditure (direct expenditure plus government transfer payments)
MS Money supply (M1)
PG Price of gold, with inverse denoted as PG^{-1}
PL Price level
PM Price of imports
PW Price of wheat
TI Trade interference due to war (such as the Continental System and the American embargo)

Hypotheses are of the form $X \rightarrow Y$ ('X causes Y, with $\partial Y/\partial X > 0$'). The exchange rate and gold price inverses, ER^{-1} and PG^{-1}, are used in lieu of *ER* and PG, respectively, where indicated for a positive derivative. Shorthand for multiple hypotheses is $W, X \rightarrow Y$ ('$W \rightarrow Y$ and $X \rightarrow Y$') and $X \rightarrow Y, Z$ ('$X \rightarrow Y$ and $X \rightarrow Z$').

Bullionist model

The bullionist chain of causation is:

$$BN \rightarrow MS \rightarrow PL \rightarrow ER^{-1}, PG \qquad (1)$$

The first relation, $BN \rightarrow MS$, reflects the bullionist correct perception of the fact that banknotes constituted the monetary base and indeed the ultimate reserve of the entire financial system during the Bank Restriction Period. There was a hierarchy of banks: the Bank of England, London private banks, and country banks. *BN* (held as reserves by the country banks and London private banks) were nonredeemable, deposits at the Bank of England (held as reserves only by the London private banks) cashable only in *BN*. The country banks – but not the London private banks – issued banknotes. Reserves of the country banks were principally deposits at the London private banks.

Strictly speaking, gold coin was a component of the monetary base, but the premium on gold bullion did not have a counterpart in the premium of gold coin (guineas and, from July 1817, sovereigns) over *BN*. There was no legal market for domestic coin in terms of paper money, Gresham's law operated, and an overwhelming proportion of the guineas and sovereigns nominally in circulation or newly minted were in fact hoarded or exported. One can defend the bullionist relation $BN \rightarrow MS$ by characterizing BN as the active component of the monetary base.

For the bullionists (and antibullionists), the money supply had as components BN, country banknotes, and coin. In excluding deposits from $M1$, the writers of the Bank Restriction Period were not far off the mark. First, except in London, 'deposits' generally meant time or savings deposits rather than demand deposits. Second, excluding interbank transactions, demand deposits typically were exchanged for cash rather than transferred to another account.[4]

The second relation in the chain, $MS \rightarrow PL$, pertains to the quantity theory of money. Underlying this theory is the bullionist view that the Bank of England effectively pegged the market interest rate at 5%, by standing ready to discount all 'good' commercial bills at that rate. This became the mainstream view of historians. The implication is that the monetary base is perfectly elastic at the constant discount rate of 5%: a powerful impetus to the quantity theory.

There is good reason for the bullionist and mainstream view; for the usury laws set a 5% limit on annual interest on bills of exchange, and the discount rate of the Bank of England was fixed at this rate. While bill brokers could charge a commission and private banks could require a minimum balance (thus circumventing the usury laws), the Bank of England did not use such devices. For these three reasons, the *market* discount rate (for good bills, those eligible for Bank of England discounting) did not exceed 5% during the Bank Restriction Period. In fact, only for about a year (beginning July 1817), did the market rate even fall below 5%.[5]

However, there is empirical basis for a contrary position. First, only 'good' bills – a minority of bills – were acceptable by the Bank of England. A 'good' bill bore at least two London names and had a maximum of 65 days until maturity. Also, the submitter of a bill had to be on the Bank of England's list of clients. Second, there is good evidence that the Bank of England effectively regulated discounts via a rationing system.[6] These facts act against the quantity theory but support the concept of BN as an autonomous policy variable. The chain of causation is complete with $PL \rightarrow ER^{-1}$, PG, which is the purchasing power parity theory (given the foreign price level).

Antibullionist model

The antibullionist model involves a balance-of-payments theory of the exchange rate, with demand and supply for bills of exchange, represented by the balance of payments (BP), yielding ER^{-1} and PG. The state of the harvest (HR) determines the domestic price of grain, represented by the price of wheat (PW). The exchange rate is an ingredient in the price of imports (PM), which, together with *PW*, determines *PL*. The antibullionists saw three principal determinants of *BP*: *PW*, trade interference (*TI*), and external military expenditure (*ME*). The full antibullionist causal chain, more complex than the bullionist version, is:

$$
\begin{array}{l}
\qquad\qquad \mathrm{HR} \rightarrow \mathrm{PW} \quad \rightarrow \quad \mathrm{PL} \rightarrow \mathrm{BN} \\
\qquad\qquad\qquad \swarrow \qquad\qquad\quad \nwarrow \\
\mathrm{TI, ME} \rightarrow \mathrm{BP} \rightarrow \mathrm{ER}^{-1}, \mathrm{PG} \rightarrow \mathrm{PM}
\end{array} \tag{2}
$$

In emphasizing the price of wheat, the antibullionists recognized the highly agrarian state of the British economy, notwithstanding the industrial revolution in progress. Crafts (1985, p. 15) calculates that between 37.0 and 41.7% of the labor force was agricultural in 1801–1803. Deane and Cole (1969, p. 166) estimate that agriculture, forestry, and fishing accounted for 32.5, 35.7, and 26.1% of national income in 1801, 1811, and 1821. The antibullionist emphasis on wartime interference with trade and on external military expenditure reflected the French Revolution and Napoleonic Wars, in which Britain was engaged for much of the Bank Restriction Period. Except for brief respites (March 1802-May 1803 and April 1814–February 1815), war was continuous throughout this period until Waterloo.

The antibullionists used the real-bills doctrine to reverse the indirect $BN \rightarrow PL$ causation of the bullionists. They accepted that the Bank of England behaved passively in its note issuance, but used the real-bills theory to demonstrate that excess issue (which would increase the price level) would be returned to the Bank of England. Then – the theory extended – only non-monetary forces could cause real income and the price level to increase and would underlie the demand for discounting to finance a higher volume of transactions, whence $PL \rightarrow BN$.

The bullionists rejected this argument as false, for ignoring the fact that the Bank of England operated without restraint on its note issue. They offered, rather, as a second-best alternative to resumption of cash payments, the *policy rule* that *BN* issuance should be oriented to the exchange rate and price of gold: $ER, PG^{-1} \rightarrow BN$.

Modern bullionist model

A 'modern bullionist' would view the monetary base (essentially *BN*) as determined by demand and supply. With supply perfectly elastic at the pegged market interest rate, *BN* is neither exogenous nor the first link in the causal chain. Rather, *BN* is proximately determined by the demand for the monetary base. One implication of the modern approach is removal of the bullionist fear that *BN* could rise without limit. While true in theory, in practice this was impossible, because *BN* was subject to the interest rate peg and so was an endogenous variable. Given not only an interest rate target, but also one that was unchanged throughout the Bank Restriction Period, the

Bank of England could not induce the private sector to hold more *BN* than permitted by demand. It is important to note, however, that this argument is different from the real-bills doctrine of the antibullionists. The modern bullionist approach is not antibullionist!

A second implication concerns revision of the formal bullionist model. With *BN* endogenous and determined by demand, ideally this demand would be proxied by the usual determinants: income and the interest rate, as well as by shocks causing shifts in the demand function. With the pertinent interest rate fixed for almost the entire period and absent continuous income data, only shock variables remain for inclusion. The obvious such variables are the very ones used as exogenous elements in the antibullionist model: *HR* (as affecting *PW*), *TI*, and *ME*. Then the 'modern bullionist model' has representation:

$$\begin{array}{l} \mathrm{HR} \rightarrow \mathrm{PW} \rightarrow \mathrm{BN} \rightarrow \mathrm{MS} \rightarrow \mathrm{PL} \rightarrow \mathrm{ER}^{-1}, \mathrm{PG} \\ \qquad\qquad\quad \nearrow \\ \qquad\quad \mathrm{TI, ME} \end{array} \tag{3}$$

Abstraction for empirical testing

For empirical testing, (1)–(3) are simplified by excluding variables for which data are unavailable or incomplete. The models remain theoretically robust. *PG* can be omitted, because it plays the same role as ER^{-1} in each system and the latter is the more pertinent variable, with *PG* of interest to the contemporary protagonists primarily as a measure of currency depreciation. Dropping MS as an intermediate variable, the bullionist system (1) reduces to

$$BN \rightarrow PL \rightarrow ER^{-1} \tag{4}$$

Omitting the non-measurable variables, *HR* and *TI*, and letting ER^{-1} represent *BP* and *PM*, the antibullionist system (2) becomes:

$$\begin{array}{l} \quad \mathrm{PW} \rightarrow \mathrm{PL} \rightarrow \mathrm{BN} \\ \qquad \searrow \quad \uparrow \\ \mathrm{ME} \rightarrow \mathrm{ER}^{-1} \end{array} \tag{5}$$

Along the same lines, the modern bullionist model (3) is now:

$$\begin{array}{l} \mathrm{PW} \rightarrow \mathrm{BN} \rightarrow \mathrm{PL} \rightarrow \mathrm{ER}^{-1} \\ \qquad\quad \nearrow \\ \qquad \mathrm{ME} \end{array} \tag{6}$$

Lessons from previous empirical testing

Results of studies

Time-series investigations of the bullionist and antibullionist theories have been performed both by contemporary writers: Ricardo (1811, pp. 114–121), Galton (1813), and Anonymous (1819); and by historians: Silberling (1923, pp. 240–243, 246; 1924, pp. 230–232), Angell (1926, pp. 477–494), Viner (1937, pp. 142–144), Morgan (1943, pp. 30–47), GRS (1953, vol. 2, p. 932), Myhrman (1976, pp. 187–189), Arnon (1990, pp. 15–16), and Nachane and Hatekar (1995) [hereinafter Nachane-Hatekar].

Ignoring findings involving PG and considering first the bullionist model, Ricardo and Galton conclude, but Anonymous rejects, $BN \rightarrow ER^{-1}$, and Nachane–Hatekar reject $MS/YR \rightarrow ER^{-1}$, where YR is real output.[7] Galton shows $BN \rightarrow PW$ (representing PL). Silberling, Morgan and Myhrman find little relationship between BN and PL; Arnon strongly supports $MS \rightarrow PL$; while Angell and Nachane–Hatekar reject $BN \rightarrow PL$ and $MS/YR \rightarrow PL$, respectively. Finally, Myhrman affirms, but Angell and Nachane–Hatekar deny, $PL \rightarrow ER^{-1}$.

In contrast to these mixed results for the bullionist model, testing of the antibullionist theories is uniformly supportive: $PW \rightarrow PL$ (Morgan and Arnon), $ER^{-1} \rightarrow PL$ (Nachane–Hatekar), $BP \rightarrow ER^{-1}$ (Morgan, GRS, and Nachane–Hatekar), $(ME + GM) \rightarrow ER^{-1}$ [where GM denotes the value of grain imports over a base level, in lieu of the pertinent causal variable, PW] (Anonymous, Silberling, Angell, and Viner), $PL \rightarrow BN$ (Angell), and $PL \rightarrow MS/YR$ (Nachane–Hatekar).[8]

Time period and frequency of observations

It is logical that the time period for testing the bullionist and antibullionist models should be within 1797–1821, the Bank Restriction Period. This provides an insufficient number of annual observations for time-series analysis. One solution, adopted by Nachane–Hatekar, is to incorporate annual observations beyond the Bank Restriction Period; they select 1802–1838 as the time period for their analysis. Because the bullionist and antibullionist models pertain strictly to a paper standard and floating exchange rate – in British monetary history descriptive only of the Bank Restriction Period, until 1914 – it is preferable to retain the 1797–1821 time period and move to higher frequency observations. Silberling and Morgan use quarterly data, a decision followed in this paper. The observation period is 2Q 1797 – 1Q 1821 – the 96 complete quarters enveloped by the Bank Restriction Period.

Multivariate empirical testing

It is natural to apply multivariate time-series analysis to the logarithmically transformed variables of models (4), (5), and (6): log(*ER*), log(*PL*), log(*BN*), log(*PW*), log(*ME*).[9] To determine the lag length of the relationship among the five variables, VARs are fitted to the variables, with a constant term and three centered seasonal dummies as exogenous variables. Testing for a lag length up to eight quarters via a modified Tiao and Box (1981, p. 807) *M*-statistic, the result is a lag length of four quarters.[10] This implies three lagged first differences for cointegration analysis, or three lags for an unrestricted VAR in first differences.

Applying the augmented Dickey–Fuller (ADF) and Phillips–Perron (PP) tests for nonstationarity, a unit root unambiguously cannot be rejected for log(*ER*), log(*PL*), log(*BN*), and log(*ME*); but the tests conflict for log(*PW*).[11] Further testing leads to rejection of the existence of cointegration vectors, and therefore to estimation of a VAR in first differences – Δlog(*ER*), Δlog(*PL*), Δlog(*BN*), Δlog(*PW*), and Δlog(*ME*) as endogenous variables, with a constant and three seasonals.[12] Granger causality testing is performed by re-estimating the VAR with the coefficients of the causal variable restricted to be zero in the equation for the affected variable and applying the system-wide likelihood ratio statistic (see, for example, Enders, 1995, p. 316). Results are shown in Table 10.1, which should be considered in conjunction with models (4), (5), and (6). Although Granger causality is not necessarily economic causality and although the variables are expressed in first differences, nevertheless the results are instructive.

Regarding bullionism, $BN \rightarrow PL$ is supported, but so is reverse causation (the antibullionist position). Rather than $PL \rightarrow ER$, the opposite is found (direct support for the antibullionist theory), though the roundabout causation $BN \rightarrow ER$ holds. Far from *BN* an autonomous variable, the $PL \rightarrow BN$ and $PW \rightarrow BN$ results justify the antibullionist real-bills doctrine. In sum, the bullionist position is subject to some serious contradictions. Regarding antibullionism, added to the positive results for *BN* are the impacts of *PW*, *ME*, and *ER* on *PL*. Only the negative findings for *PW* and *ME* on *ER* mar

Table 10.1 Granger-causality likelihood-ratio statistics

Variables	*Explanatory variables*				
	Δ*log*(*PL*)	Δ*log*(*ER*)	Δ*log*(*BN*)	Δ*log*(*ME*)	Δ*log*(*PW*)
Δlog(*PL*)	—	13.72*	11.35*	8.60**	14.87*
Δlog(*ER*)	5.41	—	8.29**	4.09	7.66
Δlog(*BN*)	11.35*	4.43	—	3.43	11.05**
Δlog(*ME*)	11.95*	1.05	1.05	—	11.95*
Δlog(*PW*)	3.76	2.76	5.41	6.38	—

*(**) Denotes rejection of no Granger causality at 1(5)% level.

support for antibullionism. The modern bullionist model receives mixed results: $PW \rightarrow BN$ provides strong support, but $ME \rightarrow BN$ does not hold.

The ordering $\Delta\log(PW) \rightarrow \Delta\log(ME) \rightarrow \Delta\log(BN) \rightarrow \Delta\log(ER) \rightarrow \Delta\log(PL)$ is used for a Cholesky decomposition to orthogonalize the VAR residuals and obtain impulse response functions and variance decompositions of the variables. The ordering follows naturally from the Granger-causality results and methodological considerations. The only variable 'caused' by every other variable is $\Delta\log(PL)$; therefore it is last in the ordering. No variable 'causes' $\Delta\log(PW)$; so it is first. Both $\Delta\log(BN)$ and $\Delta\log(ER)$ cause $\Delta\log(PL)$, but $\Delta\log(BN)$ also causes $\Delta\log(ER)$; therefore $\Delta\log(BN)$ is third and $\Delta\log(ER)$ fourth. Methodologically, $\Delta\log(PW)$ and $\Delta\log(ME)$ are exogenous variables in the antibullionist and modern bullionist systems; therefore it is logical that $\Delta\log(ME)$ be second in the ordering. Fortunately, the correlations of the VAR residuals (Table 10.2) are so low that alternative orderings do not seriously affect the innovation accounting. The only exception is $\{\Delta\log(PW), \Delta\log(PL)\}$; but their ranks in the ordering are clearly determined by the Granger-causality results.

Impulse response functions for $\Delta\log(PL)$, $\Delta\log(ER)$, and $\Delta\log(BN)$ are graphed in Figures 10.1–10.3. The response (solid line), shown for 12 quarters, is to one standard deviation of the innovation and is bounded in each direction by two standard errors (dotted lines) of the response. Shocks in $\Delta\log(PW)$ and $\Delta\log(ME)$ increase $\Delta\log(PL)$, and the $\Delta\log(ER)$ innovation reduces it (with negative dominating positive multipliers through seven quarters), all in accordance with antibullionist theory. The $\Delta\log(BN)$ innovation initially increases $\Delta\log(PL)$, pleasing to the bullionists, but then decreases it by about the same magnitude.

All innovations have a negative effect on $\Delta\log(ER)$, that is, reduce the appreciation or increase the depreciation of the pound, which is consistent with both bullionist [for $\Delta\log(BN)$ and $\Delta\log(PL)$] and antibullionist theory [for $\Delta\log(PW)$ and $\Delta\log(ME)$]. However, the positive responses in some quarters make for relatively weaker bullionist support.

The impulse response functions of $\Delta\log(BN)$ are indicative of monetary policy. A shock in $\Delta\log(ME)$ lowers $\Delta\log(BN)$, suggesting that Bank of England private discounting is reduced to stabilize the monetary base (assuming the government expenditure is financed by the Bank of England); but it is

Table 10.2 Correlation coefficients of VAR residuals

	$\Delta log(ER)$	$\Delta log(BN)$	$\Delta log(PW)$	$\Delta log(ME)$
$\Delta\log(PL)$	–0.08	0.06	0.83*	0.02
$\Delta\log(ER)$		0.06	0.02	–0.01
$\Delta\log(BN)$			0.09	–0.06
$\Delta\log(PW)$				0.08

* Denotes significantly different from zero at 1% level.

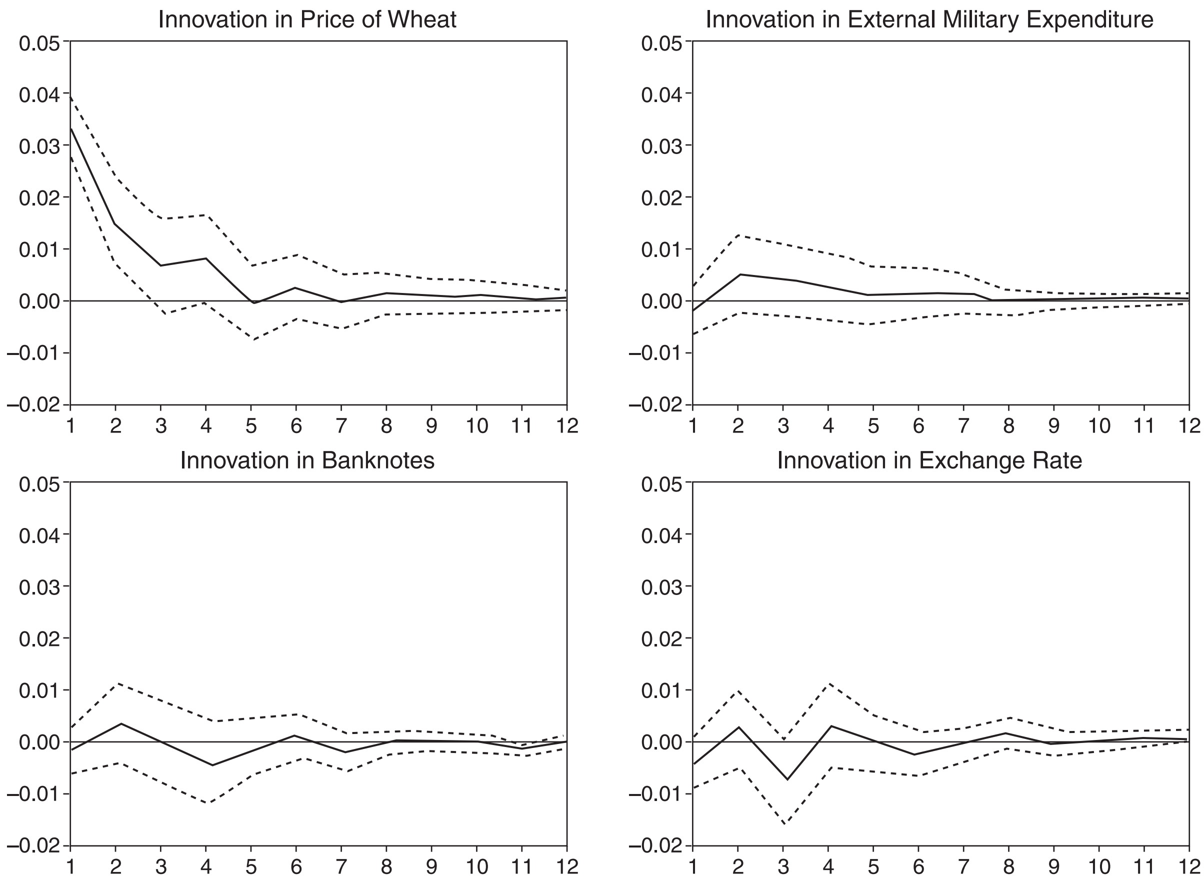

Figure 10.1 Impulse response functions of price level (first-differenced logarithms).

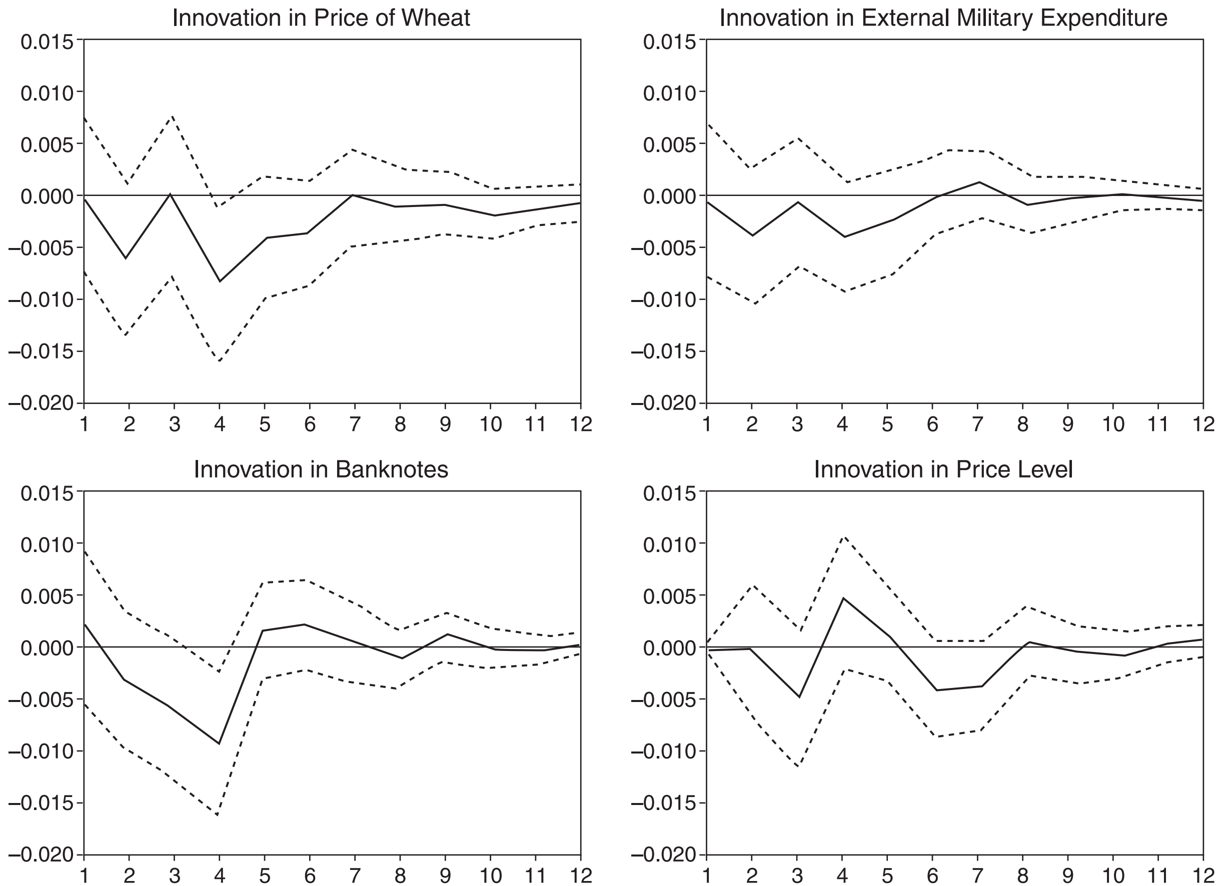

Figure 10.2 Impulse response functions of exchange rate (first-differenced logarithms).

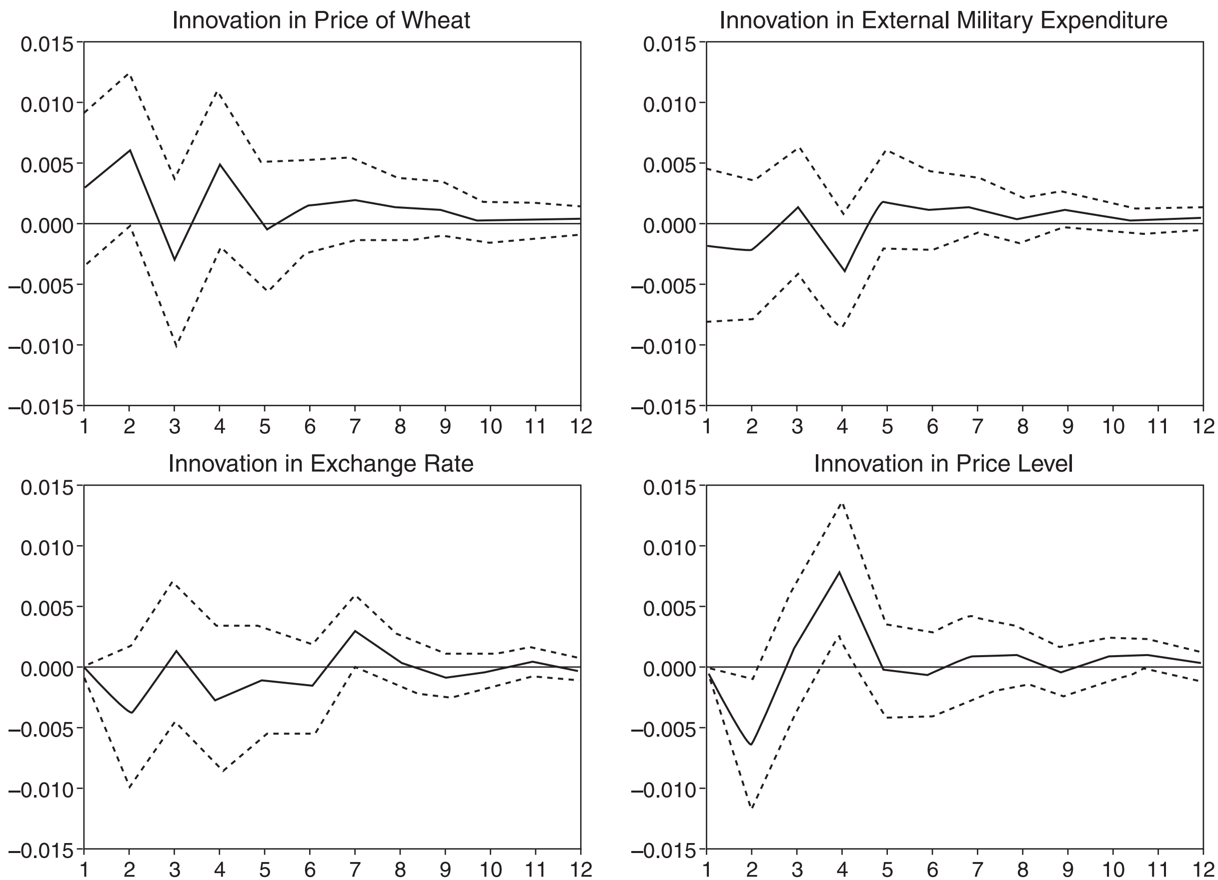

Figure 10.3 Impulse response functions of banknotes (first-differenced logarithms).

not clear which side of the controversy is thereby enhanced. However, the modern bullionist position is not supported, because it would imply a positive rather than negative impact of $\Delta\log(ME)$ on $\Delta\log(BN)$. Innovations in $\Delta\log(PW)$ and $\Delta\log(PL)$ increase $\Delta\log(BN)$ (though the latter has an initial negative effect) – supporting the antibullionist real-bills theory of accommodating monetary policy, as well as modern bullionism. Further, an innovation in $\Delta\log(ER)$, which under an activist policy (the bullionist normative rule) would increase $\Delta\log(BN)$, has the opposite effect.

Especially relevant to the bullionist debate is the variance decomposition of the three variables (Table 10.3). That for $\Delta\log(PL)$ is devastating to the bullionist case: in every period the contribution of innovations in $\Delta\log(BN)$ is the smallest among the variables and is tiny relative to the summed contributions of the antibullionist variables [$\Delta\log(ER)$, $\Delta\log(PW)$, and $\Delta\log(ME)$]. The variance decomposition of $\Delta\log(ER)$ supports the bullionists; for $\Delta\log(BN)$ makes the greatest contribution of innovations in the other variables and the sum contribution of innovations in bullionist variables [$\Delta\log(PL)$ and $\Delta\log(BN)$] exceeds that of antibullionist variables [$\Delta\log(PW)$ and $\Delta\log(ME)$] by about 45% in period 3, eventually settling to 25%. The variance decomposition of $\Delta\log(BN)$ supports the antibullionists – and, to some extent, modern bullionism – with $\Delta\log(PL)$ and $\Delta\log(PW)$ the most important other innovations. The antibullionist direction of causation $PL \rightarrow BN$ is confirmed.

Table 10.3 Variance decomposition of variables (percent)

	$\Delta log(PL)$	$\Delta log(ER)$	$\Delta log(BN)$	$\Delta log(ME)$	$\Delta log(PW)$
$\Delta\log(PL)$					
1	29.46	0.90	0.05	0.23	69.37
2	25.26	1.18	0.98	1.40	71.18
3	23.69	4.21	0.93	2.06	69.10
4	23.30	4.46	1.64	2.14	68.45
8	23.98	4.77	1.85	2.23	67.17
∞	23.97	4.80	1.89	2.25	67.09
$\Delta\log(ER)$					
1	0	99.61	0.35	0.01	0.03
2	0.01	94.87	1.20	1.10	2.82
3	1.80	90.89	3.58	1.05	2.68
4	2.83	79.71	8.70	1.86	6.91
8	4.58	76.42	8.65	2.18	8.17
∞	4.65	76.13	8.68	2.18	8.36
$\Delta\log(BN)$					
1	0	0	98.60	0.53	0.86
2	4.71	1.69	88.25	1.20	4.14
3	4.82	1.82	86.91	1.21	5.24
4	9.68	2.35	78.58	2.88	6.51
8	9.62	3.48	76.94	3.14	6.81
∞	9.72	3.57	76.69	3.16	6.86

Rows may not sum to 100.00, due to rounding.

Concluding comments

Existing empirical investigations of the bullionist experience provide mixed results for the bullionist position but uniform fundamental support for the antibullionist side. Working from general models of each side of the debate, and using multivariate time-series analysis and superior data (see Appendix A), the findings of this paper are less extreme for the antibullionist model. However, the evidence remains preponderantly in favor of the antibullionist position. There is also support for a 'modern bullionist model', similar to the antibullionist system in the endogenous quality of the monetary base (*BN*). The support might be stronger given better data availability; but this would also enhance the evidence in favor of the antibullionist model. The dichotomy between bullionism and antibullionism is reduced if contemporary bullionism is replaced with modern bullionism.

The importance of the bullionist debate is shown by its recurrence throughout monetary history when a paper currency and floating exchange rate interrupt or replace a metallic standard: Sweden in 1745–1777, France in 1788–1797, Ireland along with England in 1797–1821, and European countries after World War I.

Monetarism sees its origin in the bullionist model; and the antibullionist approach to the exchange rate (a flow theory) and monetary policy (passive, and accommodating to the price level) has gone out of fashion. It may be humbling to the macroeconomist that these theoretical developments are contravened by the preponderance of empirical results for the Bank Restriction Period.

Acknowledgements

The original article benefited from excellent referee comments.

Appendix A

Exchange rate (ER)

The exchange rate on Hamburg, the leading financial center of Europe during the Napoleonic Wars, is the only continuous London exchange rate series (apart from Lisbon, of lesser importance in trade and finance and of lower interest because also subject to a paper standard and floating rate). The exchange market instrument was the bill of exchange drawn on Hamburg and traded in London. The bill was denominated in schillings and grotes, Flemish banco. The 'Flemish' (vlamische, 'vls.') designation pertains to the Hamburg unit of account (1 pfund = 20 schilings-vls. = 240 grotes) that emanated from Antwerp. 'Banco' refers to 'bank money', transferable deposits at the Bank of Hamburg payable in silver bullion at a constant value, as distinct from various coined money that floated in value with respect to banco.[13] The schilling-pound exchange rate was given by the bill's schilling face-value/pound market-price ratio.

ER is taken from weekly tabulations in *Resumption Report (1819*, pp. 336–354) for 1797–1818 and *Bank Charter Report (1832*, pp. 98–100) for 1819–1821. The data are converted to schillings per pound and averaged quarterly; so generally there are 13 equally spaced underlying observations per quarter.[14]

Price level (PL)

The GRS (1953, vol. 1, p. 468) monthly index number of the price of domestic and imported commodities is rebased to 1796 = 100 and averaged quarterly to obtain PL.[15]

Banknotes (BN)

A quarterly series for *BN*, in millions of pounds, is obtained as the average of weekly values, from *Bank Charter Report* (1832, pp. 74–75).[16]

Price of wheat (PW)

A quarterly series is taken from Morgan (1943, p. 36) and rebased to 1796 = 100.

External military expenditure (ME)

Silberling's (1924, p. 227) annual *ME* series is the sum of (1) government expenditure on British armies in Europe and (2) government remittances abroad (subsidies, loans, and payments to foreign states and diplomatic agents). No one has attempted to improve Silberling's first component series; but the second has been superseded by the compilation of Sherwig (1969, pp. 365–368). This Chapter makes improvements to the Sherwig data.[17] The resulting series, expressed in millions of pounds, replaces (2) in Silberling's ME series. Log–linear interpolation is used to convert the series to quarterly frequency.[18]

Military expenditure abroad is so fundamental a driving force in the anti-bullionist model that the variable is included in synthetic quarterly form. The underlying assumption is that the known annual expenditures enter the quarterly balance of payments in an exponentially smooth fashion (consistent with transforming all variables into logarithms for empirical testing).

Notes

* Originally published in *International Journal of Finance and Economics*, vol. 5, no. 3 (July 2000), pp. 197–209. Copyright 2000. © John Wiley & Sons Limited. Reproduced with permission.

1 See Clapham (1945, pp. 8–9, 50), Feavearyear (1963, p. 215), and Officer (1996, p. 39).

2 Excellent histories of the Bank Restriction Period are Acworth (1925, pp. 69–114), Canaan (1925, pp. vii–xxxiv), Viner (1937, pp. 122–124, 171–174), Morgan (1943, pp. 23–48), Clapham (1945, pp. 1–74), Hawtrey (1950, pp. 268–292), Gayer *et al.* [GRS] (1953, vol. 1, pp. 47–53, 76–81, 103–107, 131–135, 159–165), Feavearyear (1963, pp. 173–224), and Deane (1979, pp. 183–194).

3 Useful surveys are Angell (1926, pp. 40–65), Viner (1937, pp. 119–160), Fetter (1965, pp. 26–54), Einzig (1970, pp. 202–207, 225–226), and Perlman (1986). The prominent bullionists were Walter Boyd, Francis Horner, William Huskisson, Peter King, Thomas Malthus, David Ricardo, Henry Thornton, and John Wheatley; while the leading antibullionists were Henry Boase, Charles Bosanquet, John Herries, George Rose, Thomas Smith, Coutts Trotter, and Nicholas Vansittart. Rare is the post-Restriction author, such as Acworth (1925, pp. 70–105), who supports the bullionist position; but many later writers follow the antibullionist line, for example: Tooke (1838, pp. 36–41, 156–170), Morgan (1943, p. 47), Clapham (1945, pp. 10–69), GRS (1953, vol. 1, pp. 59–163), and Deane (1979, p. 194). Some historians take an eclectic view, combining the bullionist and antibullionist positions: Pressnell (1956, pp. 448–449, 463–466), Feavearyear (1963, pp. 191–223), Einzig (1970, pp. 188–190, 225–226), Duffy (1982), and Vilar (1991, pp. 311–313).

4 In his estimates of M1, Cameron (1967, pp. 42–45) includes only 1/4 of total deposits in 1800–1801, 3/7 in 1811, and 5/7 in 1821; the remainder deemed to enter *M*2. The proportions should be even less, to eliminate interbank deposits. Britain's monetary system in the Bank Restriction Period is discussed in Morgan (1943, pp. 2–17), Fetter (1965), Coppieters (1955, pp. 28–55), Pressnell (1956, pp. 136–180, 190–207), and Cameron (1967, pp. 18–27, 49–51, 67–72).

5 On the usury laws and the market interest rate, see Tooke (1838, p. 159), King (1936, pp. 12, 27–29), Morgan (1943, p. 43), Clapham (1945, pp. 15, 61–62), Ashton (1959, p. 175), Homer and Sylla (1991, pp. 157, 163–164, 187, 205–206), and Duffy (1982, p. 79).

6 For the mainstream view, see Tooke (1838, p. 159), Morgan (1943, p. 47), Clapham (1945, p. 15), and Deane (1979, pp. 193–194). The contrary view is presented by Acworth (1925, pp. 145–146) and Duffy (1982).

7 Nachane–Hatekar are alone in measuring the money supply in ratio to output, which is antithetical to the literature. They measure output by the much-discredited Hoffman index of industrial production.

8 BP is proxied by the balance of trade; but there are two problems. First, the observation for 1813 is lacking, because of the records destroyed in the London Customs House fire of 1814. Second, the BP variable that fundamentally moves the exchange rate surely includes also services, income flows, transfer payments, and long-term capital movements. Only Morgan extends the *BP* proxy, by including Silberling's series on government remittances abroad (see Appendix A). It is arguable that the balance of trade is too limited a measure and that it is better to exclude the *BP* variable, a position taken in this paper.

9 Existing empirical investigations of the bullionist controversy are bivariate in nature; and, of the researchers, only Nachane – Hatekar use modern time-series analysis. They employ unit-root and cointegration tests, and then apply weak-exogenity and Granger-causality testing separately to the cointegrated and non-cointegrated cases. However, their testing remains bivariate. Also, their *BP* variable is excluded from cointegration analysis, because it is found to be *1(0)*, whereas the other variables are *I*(1). This is an outcome of bivariate modeling. Another consequence is the failure to perform innovation accounting.

10 The correction term 1/2 is removed from the *M*-statistic, to accord with the gener-

ally accepted measurement of the likelihood ratio. VARs are estimated for lag lengths 0–8 quarters for the sample 2Q 1799–1Q 1821. The *M*-statistic is nonsignificant for lag lengths 5–8 but significant at the 1% level for lag length 4. The VARs are refitted for lag lengths 0–4 for 2Q 1798–1Q 1821, with the same result for lag length 4.

11 Based on graphs of the variables, a constant is always included, and a trend is alternatively included and excluded for log(*BN*). Lag lengths 4, 8, 12 are applied to the ADF test and a truncation lag of 3 (from the Newey – West correction) for the *PP* test. Under a 5% level of significance, the PP test cannot reject a unit root for log(*PW*), while the ADF test rejects for lags 4 and 8 but not for lag 12.

12 With intercepts in the cointegrating equations and a 5% level of significance, the Johansen λ-trace test indicates two cointegrating vectors (CVs) but the λ-max test none. Estimating a VEC model, each CV exhibits a correlogram with substantial low-order autocorrelation and the *Q* test for white noise fails miserably. Similarly, in bivariate testing of four cases, Nachane – Hatekar find only one cointegration relationship. One can surmise that the upheavals of war and industrial revolution inhibited a long-term equilibrium relationship among nonstationary variables.

13 Information on the Hamburg monetary standard is in *Bullion Report* (1810, pp. 65, 73–75), Kelly (1811, 1821), Waterston (1847, p. 357), and McCusker (1978, pp. 62–63).

14 Silberling (1924, p. 231) regards the price of Spanish silver dollars in the London market 'as the most trustworthy single index, not only of the price of specie, but of the general drift of foreign exchange movements'. However, this statement is unacceptable, because (1) Britain was on a suspended gold, not silver, standard; and (2) the Spanish dollar was not equivalent to bank money and indeed was not even a circulating coin in Hamburg. Nachane – Hatekar use the exchange rate on Paris to represent *ER*. Their choice is unfortunate, because (1) there are no quotations on Paris during the Bank Restriction Period until April 1802, wherefore Nachane – Hatekar lose 5 years of good observation; (2) the exchange on Paris was not representative during wartime, especially during Napoleon's Continental System (1806–1812); (3) the Nachane–Hatekar data source (Tooke and Newmarch's *History of Prices*) has only two daily observations per year – a poor underlay of an annual series of a floating exchange rate.

15 Incomprehensibly, Nachane–Hatekar use Silberling's (1923, pp. 223–233) obsolete price index, which has serious limitations, documented by GRS (1953, vol. 1, pp. 463–483).

16 Nachane–Hatekar use *BN* plus deposits at the Bank of England (*DB*), representing the money supply (*MS*). The evidence is that a superior proxy would be *BN* itself. First, only deposits of the non-bank public constituted part of *M*1, and they were a small part of the total during the Bank Restriction Period. Annual data for 1807–1821 (*Bank Charter Report*, 1832, pp. 35, 41) show that government deposits alone composed 72–89% of deposits at the Bank, and there were also London-private-bank deposits. Second, the money-supply estimates of Cameron (1967, pp. 42–46) made for three dates during the Bank Restriction Period show that $M1/(BN + DB)$ is a more-variable ratio than $M1/BN$, with a coefficient of variation of 8.36 versus 3.23. Also, Nachane–Hatekar use data that are annual averages of only two daily observations.

17 Sherwig places all subsidies to Sicily between 1804 and 1807 in the final year; instead, they are allocated equally to the four years. He excludes loans to the House of Orange and French Bourbons in 1813–1814 (see Clapham, 1917, p. 98; Silberling, 1924, p. 225); these loans are included here. He omits the Russian–Dutch loan contracted in 1815; the payment was an annual flow, included in 1816–1822 (see *Account*, 1854, p. 470; *Return*, 1900, p. 260; Clapham, 1917, pp. 499–500).

For the purpose of quarterly interpolation, the annual series is obtained for 1796–1822.

18 The annual series, A_t, is transformed to log $A_t/4$, t = 1796, . . ., 1822. Then, centering the annual value in midyear, quarterly values are mid-quarter. Thus (1/8, 3/8) of the first-differenced series is *added* to the preceding year's value to obtain that year's (third, fourth) quarter and *subtracted* from the current year's value to yield this year's (second, first) quarter, with the resulting quarterly series denoted as B_{it}, i = 1, . . ., 4. To obtain a quarterly series consistent with the log-linear interpolation process and summing annually to A_t, an annual series $k(t)$ is constructed by solving the following equation for $K(t)$: $\Sigma\ \exp(B_{it})^{K(t)} = A(t)$. The quarterly variable corresponding to ME is $\exp(B_{it})^{k(t)}$; its logarithm, corresponding to $\log(ME)$, is $k(t)B_{it}$. A listing of ME and the other variables is presented in Table 13.3.

References

Acworth, A. W. (1925) *Financial Reconstruction in England 1815–1822*. King: London.

Account 1854. An Account of all Sums of Money Paid or Advances by Way of Loan, Subsidy, or Otherwise, to any Foreign State. House of Commons, Paper No. 466, vol. 39.

Angell, J. W. (1926) *The Theory of International Prices*. Harvard University Press: Cambridge, MA.

Anonymous (1819) Two tables showing the rates of exchange on Hamburgh compared with the amount of bank notes, and the price of gold, and with the foreign expenditure, and the value of grain imported from the year 1793 to 1819. *Pamphleteer* **15**: 281–286.

Arnon, A. (1990) What Thomas Tooke (and Ricardo) could have known had they constructed price indices. In *Keynes, Macroeconomics and Method*, D. E. Moggeridge (ed.). Edward Elgar: Aldershot; 1–19.

Ashton, T. S. (1959) *Economic Fluctuations in England 1700–1800*. Clarendon Press: Oxford.

Bank Charter Report 1832. Report from the Committee of Secrecy on the Bank of England Charter. House of Commons, Reports from Committees, vol. 6.

Bullion Report 1810. Report on the High Price of Gold Bullion. House of Commons, Reports from Committees, vol. 3.

Cameron, R. (1967) *Banking in the Early Stages of Industrialization*. Oxford University Press: New York.

Canaan E. 1925. *The Paper Pound of 1797–1821*. King: London.

Clapham, J. H. (1917) Loans and subsidies in time of war 1793–1914. *Economic Journal* **27**: 495–501.

Clapham, J. H. (1945) *The Bank of England*, vol. 2. Cambridge University Press: Cambridge.

Coppieters, E. (1955) *English Bank Note Circulation 1694–1954*. Martinus Nijhoff: The Hague.

Crafts, N. F. R. (1985) *British Economic Growth during the Industrial Revolution*. Clarendon Press: Oxford.

Deane, P. (1979) *The First Industrial Revolution*. Cambridge University Press: Cambridge.

Deane, P. and W. A. Cole (1969) *British Economic Growth 1688–1959* (2nd edn). Cambridge University Press: Cambridge.

Duffy, I. P. H. (1982) The discount policy of the Bank of England during the suspension of cash payments, 1797–1821. *Economic History Review* **35**: 67–82.

Einzig, P. (1970) *The History of Foreign Exchange* (2nd edn). Macmillan: London.

Enders, W. (1995) *Applied Econometric Time Series*. Wiley: New York.

Feavearyear, A. (1963) *The Pound Sterling* (2nd edn). Clarendon Press: Oxford.

Fetter, F. W. (1965) *Development of British Monetary Orthodoxy 1797–1875*. Harvard University Press: Cambridge, MA.

Galton, S. T. (1813) *A Chart, Exhibiting the Relation Between the Amount of Bank of England Notes in Circulation, the Rate of Foreign Exchanges, and the Price of Gold and Silver Bullion and of Wheat*. J. Johnson: London.

Gayer, A. D., W. W., Rostow and A. J. Schwartz [GRS] 1953 *The Growth and Fluctuation of the British Economy 1790–1850*, vols. 1–2. Clarendon Press: Oxford.

Hawtrey, R. G. (1950) *Currency and Credit* (4th edn). Longmans, Green: London.

Homer, S. and R. Sylla (1991) *A History of Interest Rates* (3rd edn). Rutgers University Press: New Brunswick, NJ.

Kelly, P. (1811) *The Universal Cambist*, vol. 1. Lackington: London.

Kelly, P. (1821) *The Universal Cambist*, vol. 1. Lackington: London.

King, W. T. C. (1936) *History of the London Discount Market*. Routledge: London.

McCusker, J. J. (1978) *Money and Exchange in Europe and America, 1600–1775*. University of North Carolina Press: Chapel Hill.

Morgan, E. V. (1943) *The Theory and Practice of Central Banking 1797–1913*. Cambridge University Press: Cambridge.

Myhrman, J. (1976) Experiences of flexible exchange rates in earlier periods: theories, evidence, and a new view. *Scandinavian Journal of Economics* **76:** 169–196.

Nachane, D. M. and N. R. Hatekar (1995) The bullionist controversy: an empirical reappraisal. *The Manchester School* **63:** 412–425.

Officer, L. H. (1996) *Between the Dollar – Sterling Gold Points: Exchange Rates, Parities and Market Behavior, 1791–1931*. Cambridge University Press: Cambridge.

Perlman, M. (1986) The Bullionist controversy revisited. *Journal of Political Economy* **94:** 745–762.

Pressnell, L. S. (1956) *Country Banking in the Industrial Revolution*. Clarendon Press: Oxford.

Resumption Report 1819. Reports Respecting the Bank of England Resuming Cash Payments. Lord's Committee, House of Commons, Reports from Committees, vol. 3.

Return 1900. Loans and Advances to Foreign States Return (1) of Advances by Way of Loan; (2) of Payments by Way of Subsidy to Foreign States; and (3) of Loans Raised by Foreign States of Which the Interest or Capital Has Been Guaranteed by Act of Parliament, since 1792. House of Commons, Paper No. 180, vol. 47, pp. 257–261.

Ricardo, D. (1951) The high price of bullion (1811). In *The Works and Correspondence of David Ricardo*, vol. 3, P. Sraffa (ed.). Cambridge University Press: Cambridge.

Sherwig, J. M. (1969) *Guineas and Gunpowder*. Harvard University Press: Cambridge, MA.

Silberling, N. J. (1923) British prices and business cycles, 1779–1850. *Review of Economic Statistics* 5 (Suppl.): 223–261.

Silberling, N. J. (1924) Financial and monetary policy of Great Britain during the Napoleonic Wars. *Quarterly Journal of Economics* **38:** 214–233.

Tiao, G. C. and G. E. P. Box (1981) Modeling multiple time series with applications. *Journal of the American Statistical Association* **76:** 802–816.

Tooke, T. (1838) *A History of Prices*, vol. 1. Longman, Orme, Brown, Green, and Longmans: London.
Vilar, P. (1991) *A History of Gold and Money*. Verso: London.
Viner, J. (1937) *Studies in the Theory of International Trade*. Harper: New York.
Waterston, W. (1847) *A Cyclopedia of Commerce*. Bohn: London.

11 The U.S. Specie Standard, 1792–1932

Some monetarist arithmetic*

Critical to research on the monetary history of the United States is availability of a monetary base series that is consistent, complete in coverage, and continuous over a long period. It is also important to have a balance-of-payments series with these same properties. Furthermore, the balance-of-payments series should be 'monetary' in nature, reflecting the intimate relationship between the monetary base and balance of payments. Notwithstanding the pioneering research of Milton Friedman and Anna J. Schwartz, and the follow-up work of their students and others, these series do not exist. The main objective of this article is to develop these monetary base and balance-of-payments series. The series can be used for new historical explorations and also for possible amendments of hitherto unchallenged results of previous investigations. Some examples are provided in the article, and the series are tabulated for further use by researchers.[1]

When the First Bank of the United States opened for business on December 12, 1791, the United States was effectively on a specie standard, based predominantly on the Spanish silver dollar. The Mint Act of 1786 established a bimetallic standard with domestically produced coin, but this act had not been put into effect. The specie standard was formalized into legal bimetallism (Mint Act of April 2, 1792) and then gold monometallism (Act of June 22, 1874), and it remained the norm for the country until March 6, 1933, when President Roosevelt prohibited banks from paying out gold. Successive congressional and presidential action over the next 10 months eliminated both the specie standard and any mechanism for a return to it. By contrast, during the period 1792–1932, deviations from a specie standard and fixed exchange rate – that is, paper standards and floating exchange rates – were temporary aberrations.

This 141-year period witnessed three episodes of central banking, two Independent Treasury Systems, the classic pre-1914 gold standard, and occasional suspensions of specie payments. The comparative macroeconomic performance of logically determined subperiods composing 1792–1932 is the subject of this study. A generalized exchange market pressure model is used, and annual data series are developed to fit the model, also to examine monetary pyramiding and price and income behavior. Foremost among these series is the monetary base.

The famed Friedman and Schwartz (hereafter, FS) (1963, 1970) series of the monetary base for 1867–1932 is adjusted in light of a somewhat different methodology and is extended back to 1789. Consideration is also given to the work of Rutner (1974), who provides a monetary base series in the FS tradition for 1833–1860, and Temin (1969), who generates a series autonomously for 1820–1857. Then the monetary balance of payments, consistent with the new monetary base, is generated for the full 1790–1932 period. The monetary base and balance-of-payments series are presented as fundamental data contributions, beyond the analysis to which they are put in this study.

The methodology of the historical monetary base is discussed in Section I. Whether or not the First and Second Banks of the United States were central banks seriously affects both the base and payments series, and this issue is considered in Section II, leading to separation of 1792–1932 into subperiods (Section III). The new monetary base series is generated in Section IV and presented in Section V. Comparisons with the FS, Rutner, and Temin series including amendments to historical findings, follow in Section VI. The monetary balance of payments is generated in Section VII. The new monetary base and balance-of-payments data, along with specially developed series of price, income, and other variables, are put to use in a comparative evaluation of the performance of central banking and other periods spanning 1792–1932 (Section VIII). Following conclusions (Section IX), an appendix provides details on data sources and construction of variables; the text is devoted purely to analysis.

I. Methodology of the historical monetary base

The importance of the monetary base is twofold. First, the money supply is the product of the money multiplier and the monetary base, with the multiplier being an explicit function of the commercial banks' reserve/deposit ratio and the nonbank public's currency/deposit ratio. This formulation is one of the great accomplishments of FS (1963, pp. 776–798), and they, followed by many imitators, use it repeatedly in their history to delineate the absolute and relative importance of the three determinants in changes in the money supply. Second, the monetary base is closely related to the monetary balance of payments, with a payments imbalance constituting the effect of international transactions on the monetary base. More generally, the monetary base and balance of payments, together with the exchange rate, combine to define exchange market pressure in the foreign exchange market.

The monetary base is composed of all assets that are actual or potential reserves for the consolidated commercial banking system. To make the definition operational, six questions must be answered:

1. Who holds the monetary base?
2. What are the assets that constitute the base?
3. For each asset separately, what is the time period for which it is included in the base?
4. What should be the dating pattern of the monetary base series?
5. In what money should the base be denominated?
6. What data should be used, what interpolative techniques for missing data points, and under what circumstances is information so poor that, for example, a legitimate asset should be omitted from, or an illegitimate holder should be included in, the base on statistical grounds?

Monetary base developers (and users) can reasonably differ on answers to each of these questions, depending on their objectives and the criteria that they use. The current study differs from the work of predecessors in making fully explicit these objectives and criteria.

The objectives are (a) to achieve consistency over a long duration, 1792–1932; (b) to consider the United States as on a virtual, if not actual, specie standard throughout the time span; and (c) to ensure compatibility with the monetary balance-of-payments measure. The criteria are (i) to apply strictly the definition of the monetary base and operate in accord with the objectives in answering questions 1 to 5 and (ii) to use all available information to maximum effectiveness in answering question 6. Administering criterion (ii) inevitably involves considerable judgment, and again reasonable researchers can differ in their decisions. The advantages of the current study over predecessors in this respect emanate from the work of FS and their students, the existence of specialized studies pertinent to the monetary base written since their time, and spreadsheet/statistical programs that were not available to FS.

II. Were the first and second banks central banks?

FS (1970) do not address the issue of whether the First and Second Banks were central banks. However, in showing data for these banks separate from state banks, they leave the question open. For the current study, the pertinent central banking criterion is whether the Banks' note circulation (and, by extension, non-Treasury deposit liabilities) served as actual or potential reserves for the state banking system and hence constituted part of the monetary base.[2] While the question has not been directly addressed for the First Bank, many have answered in the affirmative for the Second Bank.[3] Yet it would be a reasonable position that, given the controversial nature of these institutions and the long tradition of considering specie as ultimate money, the Banks' liabilities were considered just ordinary money. Fortunately, a variety of empirical evidence exists on the matter.

First, Fenstermaker (1965, p. 43) and Rutner (1974, p. 25, n. 1) note, for the First and Second Banks, respectively, that Bank notes were sometimes included with specie in the statements of state banks.[4] Second, Fenstermaker (1965, pp. 11–12, 69–76) synopsizes the entire history of the Second Bank in terms of its credit contraction/expansion with multiple effect on credit contraction/expansion of the state banking system.[5] Third, Engerman (1970, p. 726) and Rutner (1974, pp. 23–30, 121–146) show that the nonbank public considered Second Bank notes and deposits as substitutes for specie, the primary base money; and Rutner provides even stronger evidence for this treatment on the part of the state banks. Furthermore, the base money characteristic of Bank note and deposit liabilities continued many months after February 1836 (the date of replacement of the Bank's federal charter with a Pennsylvania charter) and even after the Bank's initial suspension of specie payments (May 1837 to August 1838) – by Rutner's evidence, until 'sometime in 1839,' probably with the Bank's second suspension in October.[6]

The same reasons underlying the monetary base property of Second Bank note and deposit liabilities apply to those of the First Bank, and hence the positive empirical findings for the Second Bank may be extrapolated to the First Bank. Each Bank was a balance sheet giant in comparison to contemporary state banks, and, as national institutions, each had branches in the major commercial cities of the country.[7] Each was the fiscal agent of the government and served as a major (First Bank) or sole (Second Bank – to 1833) despository of the Treasury. These circumstances generated a large and steady stream of state bank notes (and checks) to the Banks, which generally presented them regularly to the state banks for redemption in specie. These banks, in turn, could avoid specie loss by presenting the Bank with the Bank's notes and drawing down its deposits at the Bank. Therefore, Bank note and deposit assets were considered by the state banks as part of reserves.

The Banks' redemption practice was a technique of monetary control that was fostered by the conservative credit policy of the First Bank and by the conscious regulation of the state banks on the part of the Second Bank under President Nicholas Biddle. When the First or Second Bank chose not to redeem its state bank notes, it became a still greater creditor of these banks, thereby enhancing future control. Hammond (1947) argues that this regulatory power – different from modern central banking in the creditor rather than debtor status of the central bank with respect to commercial banks – was 'simpler, more direct, and perhaps more effective than those of the Federal Reserve Banks' (p. 2).[8]

The notes of the Banks were clearly superior to state bank notes. By federal charter, Bank notes were legal tender for all payments to the government. Combined with interstate banking, this gave rise to universal acceptability in the private sector – not a characteristic of state bank notes at the time. The conservative note issuance policy of the First Bank and the effectiveness of the Second Bank in reducing the deviation of domestic exchange rates from parity were additional elements in producing Bank note issue that compared favorably to the specie stock in uniformity and cost of transfer.[9]

III. Delineation of subperiods

As suggested in the introduction and consistent with the 'contingent rule gold standard' concept developed by Bordo and Kydland (1995) and Bordo and Rockoff (1996), there is a real sense in which the United States was on a metallic standard throughout 1792–1932, with deviations from paper currency convertibility deemed to be, and in fact, temporary. Nevertheless, subperiods of interest may be distinguished, primarily by identification of a monetary authority (First and Second Banks, Federal Reserve Banks, Independent Treasury) and secondarily by the longest suspension of specie payments (greenback period) and the 'classic' gold standard that followed.[10]

With the First Bank in operation from December 12, 1791, to the expiration of its charter on March 4, 1811, 1792–1810 is naturally the first period of central banking. The interregnum between the First and Second Banks is 1811–1816, a period of issuance of the first Treasury currency component of the monetary base (Treasury notes) and, beginning August 30, 1814, the first major suspension of specie payments. The Second Bank opened for business on January 7, 1817, and was treated as a central bank by the state banking system into 1839, yielding 1817–1838 as the second period of central banking.[11] Another interregnum, 1839–1846, includes paper standards (parts of 1839–1842 over much of the country) and the aborted first Independent Treasury System (July 4, 1840, to August 13, 1841).

The years 1847–1861 constitute the (second) Independent Treasury System, which began on January 1, 1847, when all payments to the Treasury were by law in specie or Treasury notes (not state bank notes). From April 1, 1847, payments from the Treasury were similarly made. Throughout this period, funds were kept within the government; banks were not used as depositories. The Act of August 5, 1861, began erosion of the policy, permitting proceeds of the first substantial Civil War loan to be deposited in state banks.

On December 30, 1861, virtually all banks ceased converting their notes and deposits into gold coin, and the Treasury suspended the right of holders of its demand notes to redeem them in gold. Resumption occurred on January 1, 1879, defining 1862–1878 as the greenback period. After the classic gold standard, 1879–1913, the third period of central banking began with the creation of the Federal Reserve System by the Act of December 23, 1913. The United States abandoned the gold standard on March 6, 1933, making 1932 the specie standard's last full year of operation.

IV. The new monetary base, 1789–1932: construction

Structure

In the FS tradition, the monetary base consists of all assets – gold or specie, nongold metallic money, (paper) currency, and deposits – that the consolidated private banking system can use as reserves either actually (these assets held by banks) or potentially (these assets held by the public). By definition, assets in (domestic) circulation are the sum of assets held by the banks and by the public. The monetary base is provided by 'outside' agents, and increases or decreases in components of the base occur via transactions of the 'inside' entities (the banks and public) with the outside. The outside agents are (1) the foreign sector (affecting the specie stock via international transactions), (2) the nonmonetary sector (altering the specie stock via production of bullion and consumption of bullion or coin), (3) the Treasury (producing nongold metallic money and paper currency but reducing the base by using specie as backing for issued currency), and (4) the central bank (providing paper currency and deposits, using specie as reserves for same).[12] Also incorporated are gold certificates (circulating warehouse receipts for gold deposits at the Treasury), lost currency, foreign-held currency, and nonunitary specie price of currency.

As the supply of base money (BASEs), the monetary base is the sum of the net contributions of specie, the Treasury, and the central bank. The contribution of specie is the amount of specie in the country (commonly called the 'specie stock') *minus* lost gold certificates. The gross contribution of the Treasury is its currency (excluding gold certificates) in official circulation *minus* lost currency *plus* nongold coin in circulation.[13] For the Treasury net contribution, there are two deductions: Treasury net specie (Treasury gross specie *less* Treasury gold held against gold certificates) and Treasury currency held by foreigners. Treasury gold held against gold certificates equals these certificates in official circulation: the sum of certificates in circulation and certificates lost. The gross contribution of the central bank is its currency in official circulation *minus* lost currency *plus* non-Treasury domestic deposit liabilities. The central bank net contribution is obtained by subtracting its specie and its currency held by foreigners.

Monetary variables are expressed in millions of 'gold dollars' (incorporating 'specie dollars' prior to 1860), except that the components of the gross contribution of the Treasury and the central bank are in millions of paper dollars. To convert to gold dollars, the gross contributions are multiplied by the specie price of currency (par of unity).[14]

Comparison with other historical monetary base series

Composition of base The new monetary base centers on the *net* liabilities (fiduciary contributions to the base) of the authorities, which measures the Treasury and central bank contributions given the specie stock. There are no precedents for this partitioning of the historical monetary base. The usual breakdown of the historical base focuses on the *gross* liabilities of the *combined* authorities; the specie stock is replaced by specie in circulation (specie stock *less* Treasury and central bank specie), while Treasury and central bank currency are combined. This composition – found in FS (1963, pp. 130, 179, 704–722, 735–744) and Rutner (1974, pp. 151–183) as well as in Kindahl

(1961, p. 40) – minimizes the role of specie and does not delineate the contributions of the respective authorities to the base. However, the monetary base *aggregate* is not affected by these alternative partitions.

Classification of gold certificates Circulation of gold certificates (first issued in nontrivial amount in 1866) is subsumed in the gold stock and therefore in the contribution of that stock to the monetary base. This placement is in accord with the net liabilities format and enhances the role of specie relative to the Treasury. It is in contrast to the FS treatment of gold certificates as currency. However, FS (1963, p. 25, n. 12) themselves provide two justifications for the former procedure: the pure warehouse receipt nature of the certificates and (during the greenback period) the market's refusal to recognize a premium on the certificates below that for gold itself. Again, the monetary base aggregate is invariant to where gold certificates are placed.

Dating pattern of series Uniform end-of-year dating is adopted, for consistency over the 1789–1932 time span and for compatibility with the monetary balance-of-payments series.[15] FS provide end-of-year figures only from 1907, while Rutner has 6 years that lack this dating, but their objective is rather to maximize the frequency of observations subject to a given level of data reliability. Temin's series pertains to the end of the fiscal (rather than calendar) year, because that is the timing of the flow data underlying his series.

Definition of the public Temin includes both the Treasury and the Second Bank in the public. The result is that the monetary base reduces to the specie stock. Because the Treasury did create money during the antebellum period (recognized but not emphasized by Temin), which money was used as bank reserves, Rutner is justified in treating the Treasury as an outside agent. Also, Rutner's decision to classify the Second Bank as a central bank was supported in Section II. Therefore, it is reasonable to follow Rutner in rejecting Temin's additions to the public.

The FS monetary base *includes* not only Treasury and Federal Reserve currency held by the domestic public and banks but also such currency held by the foreign public and banks (FS, 1963, p. 778; 1970, pp. 58–60). However, the FS base *excludes* U.S.-issued gold and silver coin held by foreigners. While Garber (1986) is correct in observing this inconsistency in the definition of the public, FS are simply following official data on currency and coin in circulation. It is the reporting of currency data by issuers rather than holders of money that leads to the inconsistency in the FS base. Indeed, FS note that 'in principle' and 'ideally defined,' foreign-held dollars should be excluded from the base. The FS (and Garber) ideal is followed in the current study, because data do exist to exclude foreign-held dollars from the base.

Denomination of base FS (1963) sum gold-dollar-denominated and paper-dollar-denominated components of the monetary base during the greenback period. They are well aware that this arithmetic is analogous to adding apples and oranges: 'Treating one greenback dollar as equal to one gold dollar . . . [is], strictly speaking, meaningless: it is like adding current Canadian or Hong Kong dollars to U.S. dollars on a one-to-one basis' (pp. 27–28). The same issue arises during May 1837 to August 1838, when the Second Bank suspended specie payments and its notes depreciated in terms of gold. The depreciation of Second Bank money is ignored by Rutner; but FS justify their simple summation of gold and depreciated dollars on two grounds: '[It] is done . . . in every other summary of monetary statistics for the greenback period we know of' (p. 28), and the necessary correction declines over time with the decrease in the gold premium.

Because the United States is considered in essence to be on a specie standard throughout 1789–1932, and because consistency over time is desired, the new monetary base is uniformly expressed in gold (or specie) dollars. This is done via deflation of base components that traded at a discount in terms of gold during periods of paper currency depreciation.[16] The specie price of currency for the central bank is nonunity only for 1837, and that for the Treasury is nonunity only for the greenback period. There was no central bank during the greenback period, and the depreciation of Second Bank liabilities during 1837–1838 did not affect the par value of Treasury currency.

Attention to lost currency Official currency in circulation, used in the FS base, includes 'currency irretrievably lost, destroyed, in collections, or otherwise so disposed as never to be presented for redemption' (Laurent, 1974, p. 213, n. 1); such 'lost currency' is deducted in constructing the new monetary base. FS (1963, pp. 442–443, n. 20) are aware of the issue and estimate the loss for national bank notes at about 0.1% per year, but they do not adjust their monetary base for lost currency. It may be that they judged the correction to be quantitatively unimportant based on their finding for national bank notes, or perhaps they did not see how to estimate the deduction for other forms of currency.

Treatment of state bank notes State bank notes, included in the FS base to mid-1878 (see FS, 1963, pp. 722, 724, 808), are clearly not high-powered money, are removed from the FS base by Joines (1985, p. 348), and are not a component of the Rutner base. They are excluded from the new monetary base. FS neglect to make this correction as well, probably because they deemed it to be of minor quantitative importance.

Treatment of national bank notes FS (1963, pp. 20–23, 50, 780–782) include national bank notes in the monetary base – reasonably because this currency served as a reserve for state banks and was legal tender for Treasury transactions (with exceptions). However, the current study places national bank notes in Treasury currency (and therefore in the monetary base) only from 1874, for reasons stated by FS themselves. Legally, a reserve requirement had been imposed on both notes and deposits of national banks, beginning with the first National Banking Act (February 25, 1863). Only with the Act of June 20, 1874, was the reserve requirement removed from national bank notes, while being retained on deposits. Furthermore, this act – and not the, earlier, National Banking Acts – provided for Treasury redemption of national bank notes in U.S. notes at par (based on a fund to which banks contributed 5% of their note issue, countable toward their reserves on deposits). Empirically, there was the potential, and in at least one instance (early 1873) the actuality, of national bank notes trading at a discount for U.S. notes.

Whether or not national bank notes should be included in Treasury currency and therefore in the monetary base prior to 1874, as done by FS, is a matter of judgment. On the side of inclusion is the fact that national bank notes were backed more than fully (111% of value of notes issued) by government bonds deposited with the Treasury and therefore can be construed as an indirect obligation of the government, that is, as Treasury currency at one remove. On the side of exclusion, viewed in this study as preponderant, are the existence of a reserve requirement, the absence of a redemption fund, and the trading of national bank notes at a discount in terms of greenbacks. As stated by FS (1963), '[In] the period before 1874 . . . [national bank] notes were more nearly identical with deposits than with the notes issued by the Treasury,' and '[To] treat national bank notes as part of the currency obligations of the monetary authorities . . . is of questionable appropriateness for the first few years covered by our series' (pp. 781–782).[17]

Selection of data In respect of data used, the new monetary base is closer to FS than to Rutner or Temin. The antebellum specie stock is constructed via a new technique and with substantially different data from those of Rutner and Temin. Also, Treasury gold and Treasury notes during the antebellum period have different data sources from those of Rutner. By contrast, the FS gold stock, specie stock, gold certificates, and nongold coin series are accepted and extended back to 1860. Prior to 1874, only part of Treasury currency is consistent with FS. From 1874, the entirety of Treasury currency (and of Federal Reserve liabilities, from 1914) has data compatible with FS.

The result is that the new monetary base is different from the FS, Rutner, and Temin series. Components of the net liabilities composition of the new base are discussed below. Subsequently, section V presents empirically the contributions to the base emanating from both the net liabilities and an alternative breakdown, authorities' net assets. Then section VI shows just how different the new base is from its predecessors.

Components of net liabilities composition of the monetary base

Specie stock and nongold coin Prior to 1860, data on the specie stock include both gold and silver, although by the late 1850s silver is in the form only of domestic subsidiary coinage.[18] From June 30, 1860, official specie stock series are limited to gold, consisting of domestic gold coin in circulation and gold in all forms (domestic coin, foreign coin, and bullion) in the Treasury or Federal Reserve. From that date, nongold coin (standard silver dollars, subsidiary silver coin, and minor coin) became separate official series. The specie stock for the new monetary base follows the official line – gold and silver to the end of 1859 and gold alone thereafter. The specie stock, its distribution, and nongold coin circulation agree with the corresponding FS series.

Contribution of central bank to monetary base Both the First and Second Banks issued not only banknotes (payable in specie on demand) but also postnotes (payable in specie on demand at a specified future date after issuance). For the First Bank, postnotes are included in central bank currency because (1) they were issued regularly only by the main office and in the ordinary course of business and (2) '[Total] note circulation was deliberately restricted to guard specie' (Wettereau, 1937, p. 283); there was never a question of suspending specie payments. For the Second Bank, postnotes are excluded. The Second Bank first issued postnotes in March 1837 decidedly not in the ordinary course of business, while specie payments were suspended, in an attempt to obtain specie.[19]

Contribution of Treasury to monetary base The FS composition of Treasury currency is followed in its inclusion of national bank notes (from 1874), silver certificates (receivable for all payments to the Treasury from inception in 1878, and a legal reserve for national banks by the Act of July 12, 1882), Treasury notes of 1890 (a full legal tender), U.S. notes (greenbacks, first issued in 1862, a legal tender with exception for certain payments to the Treasury), fractional currency (instituted in 1863, a substitute for subsidiary silver coin), and certain Civil War issues designated as 'other U.S. currency' in official statistics: old demand notes (payable for all public dues, made legal tender by the Act of March 17, 1862), Treasury notes of 1863, and compound interest notes (both interest bearing but legal tender on the same basis as U.S. notes).

However, Treasury currency differs from the FS concept in two respects. First, gold certificate circulation is subsumed in the gold stock, in contrast to the FS treatment of gold certificates as currency. Second, 3% certificates, issued after the Civil War, were a legal reserve for national banks and so are included in Treasury currency.[20] Also, the FS concept must be broadened in two respects for extension prior to 1867. First, postage currency, issued for nearly a year beginning July 1862 and replaced by the fractional currency, is included in Treasury currency (in fact, the two types of currency are intermixed in official statistics).[21] Second, Treasury notes, issued between 1812 and 1861, also are included in Treasury currency; interest bearing, they had the same legal tender characteristic as did Bank of United States notes, were used as bank reserves, and (in small denominations) even served as hand-to-hand currency.[22]

V. The new monetary base, 1789–1932: presentation

Net liabilities breakdown

The new monetary base for 1789–1932 is listed in Table 11.1. The contributions of the specie stock, Treasury, and central bank are presented as period

Table 11.1 Monetary base and monetary balance of payments, 1789–1932 (millions of gold dollars)

Year	Monetary base (end of year)	Balance of payments: Net specie imports		Year	Monetary base (end of year)	Balance of payments: Net specie imports	
		Direct	Indirect			Direct	Indirect
1789	9			1861	317	–5	–8
1790	16		7	1862	416	–41	–47
1791	18		2	1863	482	–74	–84
1792	22		3	1864	421	–72	–82
1793	24		1	1865	506	–59	–69
1794	25		–3	1866	452	–45	–54
1795	21		–3	1867	399	–46	–53
1796	20		–1	1868	369	–46	–51
1797	18		–2	1869	410	–24	–28
1798	22		3	1870	442	–41	–48
1799	22		1	1871	436	–50	–58
1800	22		0	1872	419	–39	–47
1801	21		0	1873	426	–25	–32
1802	18		0	1874	700	–34	–42
1803	22		2	1875	672	–38	–43
1804	22		1	1876	700	–12	–17
1805	23		1	1877	732	2	–2
1806	23		2	1878	746	2	1
1807	24		0	1879	867	39	78
1808	23		0	1880	1001	87	67
1809	25		0	1881	1113	50	63
1810	26		0	1882	1148	4	–28
1811	21		1	1883	1180	–6	23
1812	25		1	1884	1210	–1	–10
1813	28		1	1885	1202	–3	11
1814	35		1	1886	1219	5	8
1815	48		6	1887	1285	28	37
1816	33		–1	1888	1315	–13	–27
1817	38		–1	1889	1333	–28	–42
1818	35		–1	1890	1420	–38	–7
1819	30		–1	1891	1483	–35	–43
1820	29	–1	1	1892	1502	–44	–63
1821	33	–3	2	1893	1598	–46	–16
1822	32	–6	0	1894	1498	–17	–83
1823	31	1	2	1895	1441	–54	–72
1824	34	–2	–2	1896	1501	–17	43
1825	41	1	3	1897	1569	75	1
1826	38	2	1	1898	1732	77	140
1827	41	–1	–1	1899	1821	23	9

(Continued Overleaf)

Table 11.1 Continued

Year	Monetary base (end of year)	Balance of payments: Net specie imports		Year	Monetary base (end of year)	Balance of payments: Net specie imports	
		Direct	Indirect			Direct	Indirect
1828	44	2	2	1900	2025	3	28
1829	44	4	2	1901	2099	7	0
1830	48	5	2	1902	2195	–1	11
1831	56	–5	–5	1903	2309	6	19
1832	51	6	4	1904	2413	–12	–36
1833	55	6	0	1905	2505	8	7
1834	61	14	11	1906	2715	59	104
1835	76	5	4	1907	3021	68	97
1836	73	8	7	1908	3054	13	–25
1837	74	–3	–3	1909	3084	–63	–88
1838	90	9	7	1910	3161	–13	9
1839	76	13	13	1911	3238	20	24
1840	75	–1	–1	1912	3320	–10	23
1841	74	–4	–3	1913	3403	–35	–35
1842	82	6	6	1914	3386	–18	–168
1843	86	14	12	1915	3788	288	416
1844	81	–2	–2	1916	4413	516	461
1845	78	–2	–2	1917	5436	219	250
1846	90	11	9	1918	6302	–216	–219
1847	102	6	6	1919	6504	–284	–287
1848	97	–4	–4	1920	6670	–108	–125
1849	94	–1	–2	1921	5668	610	630
1850	118	–13	–19	1922	5804	155	174
1851	142	–31	–36	1923	6029	253	259
1852	165	–30	–36	1924	6340	248	270
1853	191	–29	–34	1925	6529	–39	–54
1854	202	–44	–45	1926	6481	61	67
1855	201	–47	–47	1927	6621	–57	–33
1856	205	–49	–50	1928	6599	–225	–208
1857	216	–45	–46	1929	6485	152	158
1858	244	–45	–45	1930	6678	364	367
1859	235	–57	–56	1931	7287	–239	–237
1860	250	–21	–43	1932	7673	160	148

averages in Table 11.2. The contributions of the Treasury and central bank can be interpreted as the reduction in the monetary base should the Treasury or central bank be reclassified to the private sector. In particular, treating the First and Second Banks as commercial banks would reduce the monetary base by 18 and 20%, on average. In principle, the contribution of the Treasury or the central bank can be negative, and in fact that of the Treasury is negative during 1849–1857 and 1917–1932, averaging –$813 million during 1914–1932. FS (1963, pp. 391–393, 399) consider a hypothetical policy of a Federal Reserve $1 billion open-market purchase of securities in 1930 or 1931, which would have moderated, and possibly prevented, the crises that led to the Great Depression. All the while, the Treasury was immobilizing a *greater*

Table 11.2 Contributions to monetary base: net liabilities breakdown (period means – end of year)

	Millions of gold dollars			*Percentage of monetary base*		
Period	*Specie*	*Treasury*	*Central bank*	*Specie*	*Treasury*	*Central bank*
1791–1810	18	—	4	82	—	18
1811–1816	25	7	—	82	18	—
1817–1838	38	1	9	79	1	20
1839–1846	76	5	—	94	6	—
1847–1861	185	0	—	101	–1	—
1862–1878	149	364	—	31	69	—
1879–1913	931	983	—	47	53	—
1914–1932	3708	–813	3141	61	–11	50

amount of gold; its net contribution to the base was *negative* \$1167, \$1091, and \$1359 million during 1929–1931. Treasury action to increase its monetary base contribution to *zero* was a logical alternative to Federal Reserve policy.

As expected, the share of specie in the monetary base is highest during the Independent Treasury period and lowest during the greenback period. Perhaps surprising is that only during the latter period and 1879–1913 does the Treasury make a large relative contribution to the monetary base.

Assets breakdown

The composition of the monetary base that leads directly to the balance of payments centers on the assets of the combined Treasury and central bank (with new variables measured in millions of gold dollars). The monetary base ($BASE_s$) is the sum of (1) the specie stock (SPST), (2) net foreign assets (*excluding specie*) of the Treasury and central bank (NFA), and (3) the residual contribution of the Treasury and central bank to the base (RCON):

$$BASE_S = SPST + NFA + RCON. \tag{1}$$

NFA consists of Treasury and central bank currency held by foreigners (liabilities of the authorities, therefore with negative sign) *plus* central bank net foreign assets other than foreign-held currency.

The residual contribution of the authorities (RCON) has two positive, and three negative, components. The positive terms are (i) Treasury currency in official circulation *less* currency lost *plus* nongold coin in circulation (from 1860, previously in the specie stock) and (ii) central bank currency in official circulation *less* currency lost *plus* non-Treasury domestic deposits at the central bank, with both (i) and (ii) multiplied by the pertinent specie price of currency. These two terms represent fiat currency of the Treasury, central bank domestic credit, and physical assets of these authorities (silver stock of the Treasury [from 1860] and premises of the central bank).

The negative components of RCON are lost gold certificates, Treasury net and central bank specie, and central bank net foreign assets other than foreign-held currency. Unlike in the net liabilities composition, lost gold certificates are deducted from RCON rather than from the specie stock. With the specie stock rather than specie in circulation a component of the monetary base, Treasury net and central bank specie must be subtracted from RCON. Central bank net foreign assets other than foreign-held currency are deducted for inclusion in NFA, but foreign-held Treasury and central bank currency are negative components of NFA rather than of RCON. These groupings are preparatory for balance-of-payments derivation in section VII. The asset breakdowns of the historical base provided by previous authors – FS (1963, pp. 210–212, 796–798), Cagan (1965, pp. 333–339), and Bordo (1975, p. 511)– do not separate NFA and therefore do not link to the balance of payments.[23]

Period averages for the assets composition of the monetary base are shown in Table 11.3. The relative unimportance of NFA is noteworthy. The pattern of the specie contribution versus the two other components is the same as for Table 11.2, with the relative contribution of the specie stock a maximum under the Independent Treasury System and a minimum during the greenback period. In absolute terms, the specie stock expands more than sixfold during 1879–1913 over the greenback period and again fourfold during 1914–1932 over 1879–1913. The residual contribution of the authorities is only 19 and 22% under the First and Second Banks, respectively. Certainly, the First Bank was always a conservative institution, and the Second Bank could be described as such for a good part of its federally chartered existence.

Table 11.3 Contributions to monetary base: assets breakdown (period means – end of year)

	Millions of gold dollars			*Percentage of monetary base*		
		Treasury and central bank			*Treasury and central bank*	
Period	*Specie*	*Net foreign assets*	*Residual contribution*	*Specie*	*Net foreign assets*	*Residual contribution*
1791–1810	18	0	4	82	−1	19
1811–1816	25	—	7	82	—	18
1817–1838	38	−1	11	80	−2	22
1839–1846	76	—	5	94	—	6
1847–1861	185	—	0	101	—	−1
1862–1878	149	—	364	31	—	69
1879–1913	936	−5	983	47	0	53
1914–1932	3726	−355	2665	62	−6	44

Even during the Federal Reserve period, the authorities' residual contribution is outweighed by the specie stock.

VI. The new monetary base versus predecessors

Comparison of the new monetary base with predecessor series is instructive. The Temin (1820–1857), Rutner (1833–1859), and FS (1867–1932) series are obtained on a uniform year-end basis, compatible with the new base, via linear interpolation between adjacent figures closest to year end.[24] Temin's monetary base is the specie stock. So both the new monetary base and the new specie stock are compared to it, with the three series graphed in Figure 11.1. For 1820–1857, the new specie stock averages 24% *below* the Temin series. The new base averages 12% *above* the Temin series during the period of the Second Bank (1820–1838) but 35% *below* it thereafter. The Temin series is smoother than the new base, but after 1838 it diverges sharply upward.

The new monetary base is graphed against the Rutner series in Figure 11.2. The new base averages 10% *below* the Rutner series during 1833–1850, 13% *above* it during 1851–1854 (the only years when the new base exceeds Rutner), and 17% *below* it during 1855–1859. The two series have a broadly similar pattern until 1850.

The new monetary base is uniformly less than the FS base, as shown in Figures 11.3 to 11.5. It averages 46.7% below the FS base during 1867–1873 (principally due to the exclusion of national bank notes), 2.5% below it during 1874–1897, 1.8% below it during 1898–1917 (when non-European foreign-held dollars are deducted), and 7.5% below it during 1918–1932

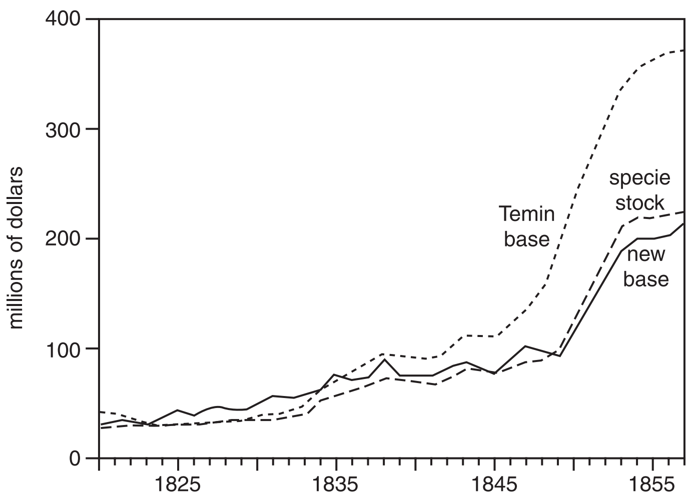

Figure 11.1 New monetary base and specie stock versus Temin monetary base, 1820–1857.

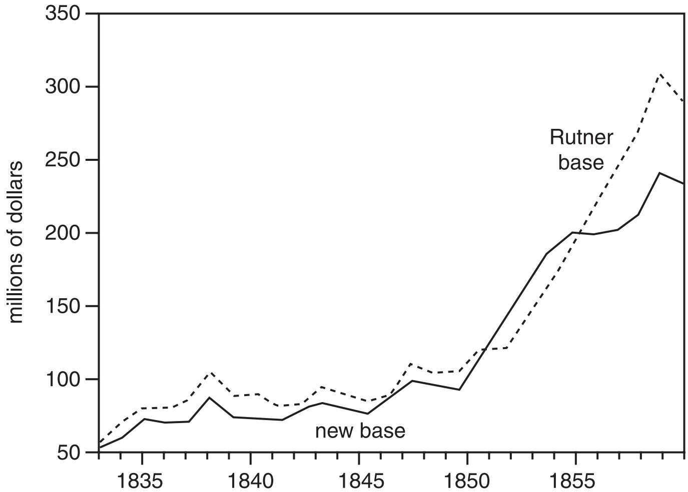

Figure 11.2 New monetary base versus Rutner monetary base, 1833–1859.

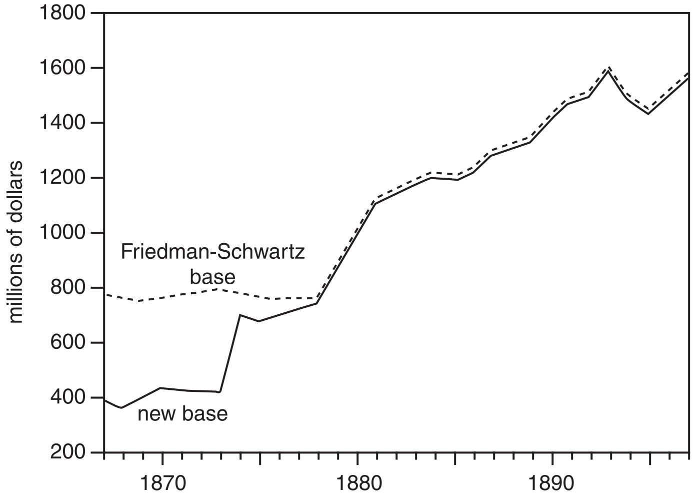

Figure 11.3 New monetary base versus Friedman–Schwartz monetary base, 1867–1897.

(when European-held dollars also are excluded). Only during 1867–1878 do the series diverge sharply. Afterward, they track each other very closely.

The new base suggests amendments to historical investigations of the determinants of the money stock. First, the new series is always below the FS base. The implication is that, for a given period or point in time, the FS series would overestimate the role of the monetary base relative to the reserve/deposit and currency/deposit ratios, compared to results using the new base.

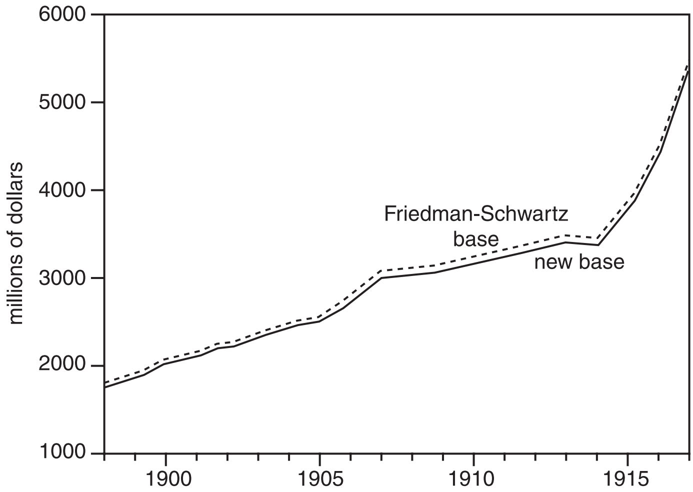

Figure 11.4 New monetary base versus Friedman–Schwartz monetary base, 1898–1917.

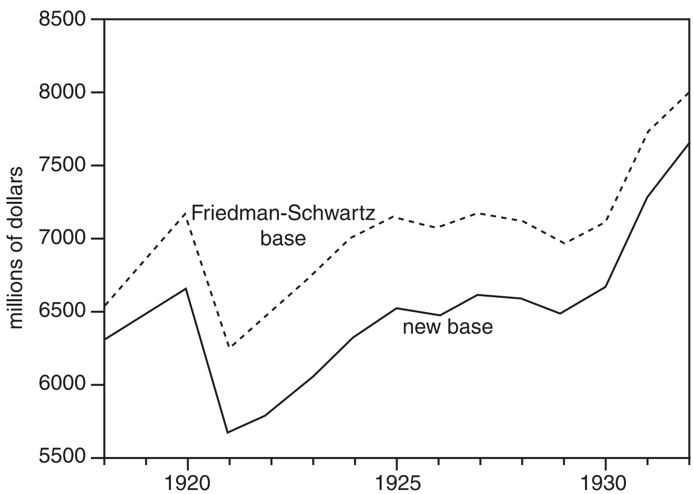

Figure 11.5 New monetary base versus Friedman–Schwartz monetary base, 1918–1932.

In this respect, the *level* of the monetary base matters – pertinent because (except for 1867–1878) in percentage changes the new series is broadly similar to its predecessors. However, researchers typically are concerned instead with *changes* in the money stock and in determinants of the money stock. Even here, the new base can make a difference. Some examples follow.

Temin (1969) observes, 'The factor leading to an expansion of the monetary stock, then, was the rise in the stock of specie. The amount of specie in the country more than doubled in the quinquennium following 1832' (p. 77). Temin shows an official specie stock series for this discussion, and the 1832–1837 increase is 184%, compared to a money supply growth of 55%. His own specie series increases by 114% (116% on an end-of-year basis). By contrast, the new monetary base increases by only 44% – *less than the money supply expansion*.

FS (1963) note 'the mild and almost horizontal movement in high-powered money' (p. 53) from January 1867 to February 1879, with their monetary base changing at an annual average rate of – 1.03% compared to 1.33% for the money supply. The new monetary base increases at an annual average rate of 4.18%, and *dominates* the other two determinants of the money supply rather than, as FS found, the reverse.[25] For July 1921 to August 1929, FS (1963) describe the 'change in high-powered money . . . [as of] minor importance for the period as a whole' (p. 275). Their figures show annual average percentage increases of 4.6% for the money supply and only 1.3% for their monetary base. The latter figure compares to 0.7% for the new base (from average 1920–1921 to average 1928–1929). In this case, using the new base *strengthens* the FS finding.

VII. The monetary balance of payments, 1790–1932

The methodology for the monetary balance of payments was developed by Kemp (1976): 'Compute the net impact of [international] transactions on the U.S. money stock . . . Of all international transactions, the only ones that affect the money stock are those that affect some component of the monetary base' (p. 10; see also Kemp, 1975a, 1975b). In this light, the existing historical balance-of-payments series – North (1960, pp. 600, 605) for 1791–1860, Simon (1960, pp. 699–705) for 1861–1900, and Bureau of the Census (hereafter 'Census,' 1975, pp. 867–868) for 1874–1932 – fall far short.

First, these series ignore net foreign assets of the Treasury and central bank, confining attention to specie transactions alone. Second, there are specific conceptual errors. The Bureau of the Census defines the balance of payments as the entire change in the gold stock, intermixing net production of gold – a purely domestic transaction – with net imports. Simon includes silver as well as gold in specie, thereby creating an inconsistency not only with official monetary data but also with the FS and new monetary bases. Third, North and Simon use official data, that measure net specie imports directly, whereas it is indicated below that an indirect computation provides the more reliable series. Fourth, North uses inconsistent data from various sources to estimate the series for 1790–1819, and his interpolation method is opaque.

Let IMP denote annual *net* specie imports and PROD annual *net* specie production, the difference between gross production and nonmonetary consumption (import of coin or bullion, production of bullion, or melting of coin that is retained as bullion or used in arts and industry *minus* nonmonetary metal melted down and recoined). Consider the equation

$$\Delta SPST = IMP + PROD. \tag{2}$$

With SPST, and therefore ΔSPST, known and of a higher order of accuracy than IMP and PROD (see Appendix), equation (2) will necessarily hold only if either IMP or PROD is estimated residually. Suppose rather that *both* variables are constructed directly, with notation IMP^{dir} and $PROD^{dir}$, and let the residual RES = ΔSPST – (IMP^{dir} + $PROD^{dir}$). Then the indirectly estimated variables are (a) IMP^{ind} = ΔSPST – $PROD^{dir}$ = IMP^{dir} + RES and (b) $PROD^{ind}$ = ΔSPST – IMP^{dir} = $PROD^{dir}$ + RES.

The monetary balance of payments (BP) is net specie imports *plus* the change in nonspecie net foreign assets of the authorities:

$$BP = IMP + \Delta NFA. \tag{3}$$

With the alternative measures of IMP, (a) BP^{dir} = IMP^{dir} + ΔNFA and (b) BP^{ind} = IMP^{ind} + ΔNFA = BP^{dir} + RES. The two balance-of-payments series are shown in Table 1, with BP^{dir} available only from 1820. While both IMP^{dir} and $PROD^{dir}$ (whence IMP^{ind}) are subject to imperfect measurement, $PROD^{ind}$ is a much more volatile series than $PROD^{dir}$, a statement not true for IMP^{ind} versus IMP^{dir}. With $PROD^{ind}$ associated with IMP^{dir}, this finding suggests that IMP^{ind} is a superior measure to IMP^{dir} (whereas no previous work has even considered using the indirect measure of specie flow) and therefore that the 'true' BP is closer to BP^{ind} than to BP^{dir}. Empirical results are shown for both BP^{ind} and BP^{dir}.

VIII. Comparative economic performance of subperiods

The eight delimited periods of 1792–1932 are compared using three sets of performance principles: monetary-oriented criteria from a generalized exchange market pressure model, monetary pyramiding ratios, and measures of price and income growth and stability.

Exchange market pressure model

The annual change in monetary base supply ($\Delta BASE_S$) is the monetary balance of payments (BP) *plus* the change in the domestic-origin component of the monetary base (ΔDOB), with the latter being the sum of net specie production (PROD) and the change in the authorities' residual contribution to the base (ΔRCON).[26] The proportionate change in $BASE_S$ is

$$\frac{\Delta BASE_S}{BASE_S} = \frac{BP}{BASE_S} + \frac{\Delta DOB}{BASE_S}$$

The most general demand-for-base function allows only for no money illusion: $BASE_D = P \cdot base_D$, where $BASE_D$ ($base_D$) is the nominal (real) demand for base money and P is the price level in specie (prior to 1860) or gold (from 1860) prices.[27] Taking the proportionate change in $BASE_D$ and imposing money market equilibrium,

$$\frac{BP}{BASE} = \frac{\Delta P}{P} + \frac{\Delta base_D}{base_D} - \frac{\Delta DOB}{BASE} \tag{4}$$

Foreign (f) money-market equilibrium:

$$\frac{\Delta BASE^f}{BASE^f} = \frac{\Delta P^f}{P^f} + \frac{\Delta base_D^f}{base_D^f}, \tag{5}$$

where the foreign price level (P^f) is in gold currency. Taking equation (4) *minus* equation (5), rearranging terms, and adding to each side the proportionate change in the exchange rate (E, the number of units of foreign currency per dollar),

$$EMP = DPP + SB + DB, \tag{6}$$

where

$$EMP = \frac{BP}{BASE} + \frac{\Delta E}{E} \qquad DPP = \frac{\Delta P}{P} - \frac{\Delta P^f}{P^f} + \frac{\Delta E}{E}$$

$$SB = \frac{\Delta BASE^f}{BASE^f} - \frac{\Delta DOB}{BASE} \qquad DB = \frac{\Delta base_D}{base_D} - \frac{\Delta base_D^f}{base_D^f}.$$

Equation (6) divides EMP (exchange market pressure in favor of the dollar) into three components: DPP (deviation from purchasing power parity in favor of the dollar), SB (monetary supply-side nominal contribution to EMP), and DB (monetary demand-side real contribution to EMP).[28] EMP has alternatives EMP^{dir} (EMP^{ind}), resulting from BP^{dir} (BP^{ind}) in its construction; similarly, it has SB^{dir} (SB^{ind}) from $PROD^{dir}$ ($PROD^{ind}$) (via ΔDOB). In equation (6), SB^{ind} (SB^{dir}) is associated with EMP^{dir} (EMP^{ind}). DB is computed from equation (6) residually, whence $DB^{dir} = DB^{ind}$.

Considering the left-hand side of equation (6), the magnitude of EMP measures external disturbance to the domestic economy, involving a change in the monetary base and/or the exchange rate, with adjustment and possibly associated costs to follow. Table 11.4 shows the period means of both algebraic and absolute values of EMP (in percentages).[29] Period efficiency varies inversely with the magnitude of either measure of EMP, but the tougher test is absolute value, as positive and negative figures reinforce rather than offset one another. Irrespective of the criterion and of whether EMP^{dir} or EMP^{ind}

Table 11.4 Exchange market pressure (period means – percentages)

	Algebraic value: Net specie imports		*Absolute value: Net specie imports*	
Period	*Direct*	*Indirect*	*Direct*	*Indirect*
1792–1810		0.83		7.09
1811–1816		7.04		8.20
1817–1838	2.06[a]	2.74	9.29[a]	6.51
1839–1846	5.26	4.89	8.69	7.76
1847–1861	–14.88	–17.00	15.93	17.89
1862–1878	–8.76	–10.26	8.86	10.27
1879–1913	0.45	0.63	1.95	2.69
1914–1932	2.08	2.07	4.42	4.87
1920–1932[b]	1.31	1.41	3.70	3.78

a Specie flow calculation indirect for 1817–1819.
b Excludes years during which London gold market was nonoperational.

is considered, the classic gold standard (1879–1913) exhibits the greatest efficiency, with the Federal Reserve period being second (absolute value measure).[30] Removing the years during which the London gold market was nonoperational (wherefore correction for paper currency depreciation could not be made) enhances performance of the Federal Reserve period but insufficiently for displacement of the primacy of 1879–1913.

From the right-hand side of equation (6), $|SB + DB| = MC$ is the absolute 'monetary component,' or the magnitude of that part of EMP contributed by the monetary supply side and demand side offsetting or reinforcing each other. The smaller the MC (as a period mean), the more efficient the period. For $MC = 0$ ($SB = -DB$), there is perfect efficiency (complete offsetting of supply and demand), but there is no maximum value of MC. For a relative measure, suppose that $|SB|$ and $|DB|$ are given to the monetary standard. Then their sum is the level against which MC is measured, whence the relative monetary component $RMC = 100 \cdot MC/(|SB| + |DB|)$, computed as a period mean. Maximum efficiency, $RMC = 0$, occurs again for $MC = 0$, but now maximum $RMC = 100$ (for $SB \cdot DB > 0$), involving reinforcement (or non-offsetting) of supply and demand contributions.

Table 11.5 shows the MC and RMC measures. The classic gold standard has maximum efficiency for MC and shares it with the 1811–1816 interregnum for RMC. The uniform superiority of 1879–1913 over central banking periods is especially noteworthy.

Monetary pyramiding ratios

The ratio BASE/SPST measures discipline, from a specie standard viewpoint, in restricting the monetary base. Under a pure specie standard, the ratio is unity. The Independent Treasury (1847–1861) and the preceding interregnum

Table 11.5 Monetary component of exchange market pressure (period means)

	Absolute: Net specie imports		*Relative: Net specie imports*	
Period	*Direct*	*Indirect*	*Direct*	*Indirect*
1792–1810		11		64
1811–1816		6		28
1817–1838	9[a]	7	56[a]	55
1839–1846	8	8	44	47
1847–1861	18	19	69	64
1862–1878	10	11	51	55
1879–1913	2	3	36	44
1914–1932	7	7	66	66

a Specie flow calculation indirect for 1817–1819.

Table 11.6 Pyramiding ratios

	Monetary base to specie stock[a]		*Money income to monetary base*[b]	
Period	*Mean*	*Coefficient of variation (percentage)*	*Mean*	*Coefficient of variation (percentage)*
1792–1810[c]	1.22	6.86	21.89	21.43
1811–1816	1.25	17.80	27.01	13.40
1817–1838	1.27	11.45	23.08	13.53
1839–1846	1.06	4.69	18.72	7.84
1847–1861	1.00	10.81	15.88	11.81
1862–1878	3.72	42.77	13.48	24.06
1879–1913	2.17	14.80	10.08	6.61
1914–1932	1.65	15.31	12.64	16.60

a Monetary base and specie stock end of year.
b Monetary base average of current and previous end of year.
c 1791–1810 for monetary base to specie stock.

come closest to the ideal ratio (zero coefficient of variation around a unitary mean), with results in Table 11.6 for the mean and coefficient of variation. As would be expected, the greenback period is least disciplined, followed by the classic gold standard (for mean ratio). Paradoxically, the flexibility of the ratio may help to explain the latter period's remarkable efficiency, in both the external economy (discussed above) and the internal economy (considered below).

Consider the further ratio $(P \cdot Y)/\overline{\text{BASE}}$, where Y is real GNP and $\overline{\text{BASE}} = (\text{BASE}_{-1} + \text{BASE})/2$. The numerator of this ratio is nominal GNP denominated in gold dollars, consistent with the expression of BASE. Then the ratio is income velocity with reference to the monetary base.[31] Period efficiency involves a low and stable velocity. Therefore, measured efficiency varies inversely with the mean and coefficient of variation of velocity. Table 11.6

shows that 1879–1913 prevails over all periods as having maximum discipline (lowest mean velocity and lowest coefficient of variation).

Price and income behavior

Period efficiency varies inversely with price instability and price volatility. Two price concepts are used: (1) 'gold-price level,' the price concept (P) in the exchange market pressure model, which corrects for depreciation of paper currency against gold or specie, thus placing paper standards on an equal footing with effective specie standards; and (2) 'paper price level,' the conventional concept, in which prices are undeflated. The indicator of price instability is mean inflation (percentage per year), computed as $100 \cdot \Delta \ln P$, while the measure of volatility is the standard deviation of trend-corrected P, $100 \cdot [\ln P - F(\ln P)]$, where F is the Hodrick–Prescott filter (smoothing parameter 100). Adjustment for trend eliminates bias in period comparisons, and for each statistic perfect efficiency involves a zero value. Results are exhibited in Table 11.7. The classic gold standard has the best performance by either criterion and for each price concept. The Federal Reserve period exhibits maximum volatility for each price level and, if truncated at 1929, maximum instability for the gold price. The 1811–1816 period has the most unstable, and the second most volatile, paper prices.

Income growth and cyclical stability are the final efficiency criteria. A fair comparison of monetary standards is enhanced by expressing growth (percentage per year) in per capita terms: $100 \cdot \Delta \ln(\mathrm{YC})$, where YC is per capita real income, but cyclical income is in overall terms and trend corrected, $100 \cdot [\ln(Y) - F(\ln Y)]$. Monetary standard performance increases with mean growth and decreases with the standard deviation of cyclical income. As

Table 11.7 Monetary-standard performance: price and income

Period	*Mean rate of growth (percentage)*			*Standard deviation[a] (detrended logarithm)*		
	Gold price	*Paper price*	*Per capita income*	*Gold price*	*Paper price*	*Income*
1792–1810	1.26	1.26	1.04	5.09	5.24	3.46
1811–1816	2.62	4.06	0.30	6.86	9.34	3.20
1817–1838	–1.48	–1.80	0.94	6.45	6.44	3.32
1839–1846	–2.63	–2.82	0.32	5.60	5.81	3.67
1847–1861	1.19	1.19	1.50	4.15	7.73	4.18
1862–1878	0.46	0.50	1.94	7.94	8.56	4.50
1879–1913	0.21	0.19	1.93	2.66	2.77	3.44
1914–1932	1.65	1.65	–0.50	9.73	9.73	7.50
1914–1929[b]	3.47	3.47	1.65	10.08	10.08	6.60

a Multiplied by 100.
b Truncated period ending with year of peak income.

shown in Table 11.7, 1879–1913 is trivially behind the greenback period in highest mean growth but substantially superior to it in income stability. Truncating the Federal Reserve period to end in 1929, the year of peak income, transforms its negative growth to only third highest, and the other central bank periods exhibit growth below even that of the Independent Treasury System.

The classic gold standard also performs well in cyclical stability, third to the 1811–1816 and Second Bank periods. Even with the 1929 truncation, the Federal Reserve period shows maximum instability, followed by the greenback period.

Comparison with existing literature

The existing literature on economic performance of historical periods is quite different from the current study.[32] The usual objective is multicountry comparison of the classic gold standard and/or comparison of that era with later periods, whereas the current study is strictly concerned with the United States and over a long time span. Previous studies ignore the greenback period, rarely consider the antebellum experience, and do not delineate pre-1914 subperiods by monetary authority. Also, the existing literature disregards exchange market pressure and monetary pyramiding, considering only price and income behavior. The only findings at all comparable to the current study are Meltzer (1986) and Miron (1989), and for truncated classic gold standard and Federal Reserve periods. Meltzer's results are unfavorable to the classic gold standard, which has greater 'risk' and 'uncertainty' for real output, whereas Miron's findings are consistent with the current study.

IX. Conclusions

Friedman and Schwartz's (1963) book is properly judged as 'surely one of the most important books in economic history, and indeed, in all of economics, written in the twentieth century' (Rockoff, 2000, p. 1). The current study builds on the FS fundamental data contribution, their monetary base series, and extends it back to 1789. The result serves as the foundation for a monetary balance-of-payments series (in two versions) over 1790–1932 – a time span during which the United States was actually or potentially on a specie standard.

The years 1792–1932 are divided into eight periods (including three central banking episodes, the Independent Treasury System, the greenback period, and the classic gold standard), and their economic performances are compared. A generalized exchange market pressure model naturally evolves from the data development, with the balance of payments deflated by the monetary base. Criteria of monetary standard efficiency, developed from this model and also from monetary pyramiding considerations, are used in addition to the usual price and income behavior. The main empirical result is the undeniable superiority of the classic gold standard (1879–1913) over central banking episodes and all other periods.

Appendix: data sources and construction of variables

A. Specie stock

1789–1859 Existing specie stock series for the antebellum period all are generated by adding a net specie flow series cumulatively to a base figure.[33] The principal problem with this technique is the incomplete nature of the official specie trade series, due to (1) smuggling, (2) unavailability of a reliable return on silver imports from Mexico, (3) absence of the requirement that overland exports to Canada and Mexico be declared, and (4) lack of data on gold and silver brought in by immigrants (until the mid-1850s).[34] It follows that a superior method involves *benchmark specie-stock estimates emanating from data on specie held by the banks (including the central bank), Treasury, and public*, without the use of flow data. An annual specie flow series then serves to interpolate between benchmark dates, with resort to linear interpolation where the series is inadequate or unavailable. The underlying assumption is that the benchmark figures are better estimates than those obtained by cumulating specie flows.

For end-of-year 1789–1806, Blodget (in Treasury *Report*, 1855, p. 51) is the specie stock (SPST) source.[35] Using only stock data, Blodget in effect counts specie held by the banks and public (FS, 1970, p. 233). Gallatin (1831, pp. 45, 49, 53–54), who was well aware of Blodget's work, provides figures for specie in banks for end-of-year 1810, 1814, 1815, 1819, and 1829. He generates the (end-of-year) 1829 specie stock explicitly as the sum of specie in banks and in the public. His technique for estimating public holdings of specie is applied here to his data for the other years, resulting in corresponding benchmark specie stock figures.

Woodbury provides specie held by banks and by the public for various dates; his basic figure is for the end of 1833, providing another benchmark estimate.[36] The final benchmark figure is for end-of-year 1860 and is the gross specie stock, constructed as the sum of the gold stock (computed below) and subsidiary silver stock (the latter 'midyear-averaged' [the average of the current and subsequent midyear] figures for 1860 and 1861, in Treasury *Report*, 1928, p. 552).

Specie stock figures between the benchmark dates remain to be determined. The interpolative flow series (*F*) for 1820–1860 is constructed as the sum of net specie imports and net specie production, where net specie production is domestic production *minus* nonmonetary consumption. This consumption involves gold or silver obtained via import of coin or bullion, domestic production of bullion, or melting of coin (but not via reworking of existing nonmonetary metal) that is retained as bullion or used in arts and industry *minus* nonmonetary metal (in jewelry or other manufacture) melted down and recoined.

Net imports of gold and silver are official data, available from 1820 and customarily used by researchers.[37] However, better production and consumption series have become available since earlier research or were ignored in previous work on the specie stock. Annual gold production for 1820–1847 is the 'middle' estimate of Martin (1976, pp. 446–447), with the total for 1792–1823 divided equally among the years. For 1848–1860, the source is Berry (1984, pp. 74, 76). Silver production is from Herfindahl (1966, pp. 323, 328–329).[38] Seaman (1852, pp. 258–260) is the source of nonmonetary consumption for 1820–1850.[39] Figures for 1851–1860 are obtained via linear interpolation between 1850 and 1880 (new gold and silver used in manufacturing and the arts, in Mint *Report*, 1921, pp. 62–63).[40] Gold consumption in 1880 is gold used in manufactures and the arts, total new material (Mint *Report*, 1921, p. 62). Silver consumption in 1880 is the product of ounces of silver used in manufactures and the arts, total new material (Mint *Report*, 1921, p. 63), and the price of silver (Census, 1975, p. 606).

To interpolate between successive benchmark estimates, years 0 and n, let $\Delta_0^n \text{SPST} = \text{SPST}_n - \text{SPST}_0$ (with the gross specie stock used in place of SPST_{1860}, which is purely gold), and note that Δ_0^n SPST is uniformly positive. Linear interpolation is applied where F is unavailable (1807–1809, 1811–1813, 1816–1818) or $\sum_{i=1}^{n} F_i$ is negative (1820–1828). This leaves $(n, 0)$ = (1833, 1829) and (1860, 1833). Intervening years $j = 1, \ldots, n-1$ are obtained as

$$\text{SPST}_j = \text{SPST}_0 + \left(\frac{\sum_{i=1}^{j} F_i}{\sum_{i=1}^{n} F_i} \right) \cdot \Delta_0^n \text{SPST}.$$[41]

1860–1932 The gold stock series constructed by FS (1963, p. 723; 1970, pp. 353–354) for 1867–1932 is replicated and extended back to 1860, with some different data sources. Published figures for 1860–1877 are only for midyear (June 30). Except for 1873–1879 in Mint *Report* (1941, p. 84), pre-1879 figures are not corrected for gold presumed lost by the Director of the Mint. Unrevised data for midyear 1860–1872 and end-of-year 1878 are in Treasury *Report* (1898, p. 109; 1928, p. 552).

The FS gold loss series for midyear 1867–1873 is the difference between uncorrected (Treasury *Report*, 1928, p. 554) and corrected (Kindahl, 1961, p. 40) gold plus gold certificates in circulation. Rounded to the nearest million, the figure equals that shown by FS (1963, p. 17) for 1867 and is readily extended to 1860 via FS's linear interpolation. Subtracting gold loss from the unrevised stock, the corrected stock series is obtained for midyear 1860–1872.

End-of-year specie stock for 1860–1877 is computed via midyear averaging. For 1878, it is the difference between the uncorrected stock and the average of the midyear 1878 and 1879 gold loss. For end-of-year 1879–1932, the source is Mint *Report* (1941, p. 84).

B. Net specie imports and production

Net specie imports (IMP^{dir}) are from section A for 1820–1859, calendar-year annualized net gold imports (same source) for 1860–1932, *minus* calendar-year annualized addition to gold exports to Canada (Simon, 1960, p. 645) for 1860–1893, *plus* change in earmarked gold (Board of Governors of the Federal Reserve System [hereafter 'Governors'], 1943, p. 536) for 1916–1932. Net specie production ($PROD^{dir}$) is production *minus* nonmonetary consumption, where production is from section A for 1820–1859, computed as described in section A for 1792–1819, the 1792–1823 annual value repeated for 1789–1791, and gold production for 1860. For 1861–1900, the source for gold production is Berry (1984, p. 78); for 1901–1932, production in fine ounces (Census, 1975, p. 606) is multiplied by price ($20.67 per fine ounce).

Nonmonetary consumption is from section A for 1820–1859 and computed as described in section A for 1811–1819. The percentage decline 1831–1821 is applied to 1801–1791 and 1811–1801 (as for 1821–1811), and linear interpolation is used for intervening values. Consumption for 1860–1932 is of gold alone. For 1860, gold consumption is estimated as the product of 1860 specie consumption and the 1880 ratio of gold consumption to gold plus silver consumption (from section A). Linear interpolation between 1860 and 1880 is used for the intervening years. Gold consumption for 1881–1932 is gold used in manufactures and the arts, total new material (Mint *Report*, 1921, p. 62; 1933, p. 30).

C. Lost and foreign-held currency

Lost currency Lost Treasury currency (including gold certificates) and Federal Reserve currency, midyear 1862–1933, is obtained as the difference between listed and loss-adjusted circulation of currency denominated up to $1000.[42] Estimated lost national bank notes are excluded during 1862–1874, via subtraction of the product of the computed lost currency and the official circulation ratio of national bank notes to the sum of national bank notes, old demand notes, U.S. notes, and gold certificates (data from Treasury *Report*, 1928, p. 554, and for old demand notes, from Treasury *Reports*, 1862–1874; the ratio is zero during 1862–1863). Midyear averaging of the resulting series yields end-of-year figures for lost currency 1862–1932.

Foreign-held currency Countries for which dollar holdings are available are those in Europe (including Britain), Cuba, the Dominican Republic, and Honduras. The initial European stock of dollars at end-of-April 1923 (Governors, 1943, p. 417) is assumed to derive from equal annual flows, beginning with a zero stock at the end of 1917.[43] The April 1923 stock is assumed to equal the measured cumulative net outflow May 1923 to June 1938; with the May-December flow included to yield end-of-year 1923 and subsequent annual net flows added to provide end-of-year 1924–1932 figures; this is Garber's (1986, pp. S150–S151) methodology.

The stock of dollars in Cuba is from Wallich (1950, pp. 320, 324–325). He provides an end-of-year series for 1920–1932. For 1912, the stock of dollars is taken as the midpoint of Wallich's range for coin plus dollar bills *minus* the midpoint of the range for coin. The dollar stock in 1912 is interpolated linearly to zero in 1897, on the assumption that dollars in Cuba reached a measurable level only with the Spanish–American War.[44] For 1918, Wallich adds a range of $10 million to $15 million to the contemporary estimate of coin plus dollar bills and interprets the contemporary author's method as dollars constituting two-thirds of the total. Taking the midpoint of Wallich's range, the arithmetic is clear for the 1918 estimate. Linear interpolation yields figures for 1913–1917 and 1919.

The data source for dollars held in the Dominican Republic and Honduras is Mint *Reports* (1917–1931). It is assumed that an April 1917 figure for the Dominican Republic applies to end-of-year 1916. In 1905, President Theodore Roosevelt imposed a customs receivership on the country, and the dollar was adopted as the standard of value. Therefore, the dollar stock of the Dominican Republic is deemed to have increased linearly from zero in 1904 to the 1916 figure. Existing end-of-year figures are 1917, 1919–1920, 1922–1923, and 1925–1930 for the Dominican Republic and 1920–1921 and 1924 for Honduras.[45] Linear interpolation between adjacent years is applied to obtain 1918, 1921, and 1924 for the Dominican Republic and 1922–1923 for Honduras. It is assumed that 1931–1932 values for the Dominican Republic are the same as the 1930 value.

As a consequence of a rise in the price of silver above 70¢ per ounce (1917–1920 according to Census, 1975, p. 606), $3 million in U.S. currency was imported into Honduras by end-of-year 1920 (Mint *Report*, 1921, pp. 154–155). This amount is allocated equally over these 4 years. The residual stock at end-of-year 1916 is an end point for linear interpolation to 1904, as for the Dominican Republic. In 1926, the government of Honduras took steps to discourage dollar circulation (Mint *Report*, 1927, p. 127). Therefore, the 1925 figure, assumed to be the same as the 1924 figure, is halved for 1926 and halved again for 1927–1932.

D. First Bank variables

Specie Specie holdings for 1792–1800 are from the Bank's (closest to) end-of-year balance sheets prepared by Wettereau (1985); for 1791, with no branches, the figure is for the Philadelphia main office alone. For 1801, 1808, and 1810, Wettereau's presentation of the November 26, 1801; February 1809; and January 15, 1811, balance sheets of Gallatin are used. For 1802, the figure of $9 million is taken, based on Gallatin's statement in November that specie holdings were more than $8 million and still increasing (Wettereau, 1937). The 'alarmingly low figure' in May 1804 and February 1805 suggests an amount of $2 million for (end of) 1804, exceeded only from 1797. By May 1806, with note circulation about $5 million, 'the specie problem was no longer acute, the supply on hand exceeding the total note circulation' (p. 283), implying holdings of $6 million for end-of-year 1806. Linear interpolation between adjacent figures is used for the remaining years.

Currency The same sources as for specie provide circulation for 1791–1801, 1808, and 1810. Figures for 1803 and 1807 are taken from House Document 27, 23rd Congress, 2nd session (hereafter 'HD27 23C 2s'). Linear interpolation between adjacent figures is used for the remaining years.

Non-Treasury deposits The same sources as for specie are used for 1791–1801, 1808, and 1810, but only total deposits are given for 1791 and 1808. To estimate non-Treasury deposits for 1791, the proportion of non-Treasury deposits for March 9, 1792, is applied. Treasury deposits at the Bank, available for 1791–1801 and 1810, are obtained for 1803–1806 from Holdsworth and Dewey (1910, p. 60) and estimated via linear interpolation of adjacent years for 1802 and 1807–1809. This permits computation of non-Treasury deposits for 1808 as a residual and for 1809 as the average of 1808 and 1810.

Assume that the modern reserve ratio, defined by the equation 'reserve ratio *equals* specie holdings *divided by* the sum of currency in circulation, non-Treasury deposits, and Treasury deposits,' was a meaningful statistic for this conservatively operated bank.[46] In 1802, specie holdings, at their highest level to that date, are in the same order of magnitude as in 1809; in 1803, they are very close to the 1800 and 1801 values; and in 1804, they are extremely low, taken as $2 million. Therefore, it is assumed that (1) the reserve ratio for 1802 is the same as that for 1809, (2) the reserve ratio for 1803 is the average of the 1800 and 1801 values, and (3) the reserve ratio for 1804 is the average for 1792–1796, the previous years when specie holdings were less than $2 million. For 1805–1807, the reserve ratio is linearly interpolated between 1804 and 1808. Non-Treasury deposits for 1802–1807 are then obtained via the reserve ratio equation.

Net foreign assets These are net assets on 'foreign transactions' account *plus* holdings of foreign bills of exchange *minus* Amsterdam loan outstanding. The source is the same as for specie.[47] Except for the Amsterdam loan, foreign assets and liabilities are listed in the Bank's balance sheets only for 1792–1795. Using information in Wettereau (1937, p. 269, n. 27), a complete series of the outstanding amount of the Amsterdam loan – a relatively large item – is constructed. It is unknown whether there were other foreign items during the years for which balance sheets are not extant.

E. Second Bank Variables

Regarding Second Bank data, Smith (1953, p. 276) is suspicious of the much-used Tyler Report (Senate Document 17, 23rd Congress, 2nd session (hereafter 'SD17 23C 2s']) and recommends using the Bank's actual returns whenever possible, the procedure followed here.

Specie Consulting the end-of-year returns printed in various congressional documents, Bank specie is obtained for 1821–1823 and 1825–1838.[48] For the remaining years, resort must be had to Tyler's data (in SD128 25C 2s). There are no data for end-of-year 1817, so linear interpolation is applied to the figures for September 1817 and February 1818.

Currency Circulation for 1817–1820, 1824, and 1832–1838 is from the same sources as for specie. For the remaining years, SD128 25C 2s is used; because the pre-1832 returns show only notes issued, notes on hand and in transit must be deducted to derive circulation.

Non-Treasury deposits Same sources as specie are used.

Net foreign assets Holdings of foreign bills (or net foreign exchange) *plus* amount due from European bankers *minus* amount due to European bankers are used. The sources are the same as for specie, except for 1837. For that year, linear interpolation is applied to the figures for December 1, 1837, and February 1, 1838 (from actual returns in SD128 25C 2s).

F. Federal Reserve variables

Specie Gold in Federal Reserve banks is from Mint *Report* (1941, p. 84).

Currency Federal Reserve notes and Federal Reserve banknotes in official circulation are from Governors (1943, pp. 409–412), the FS source.

Domestic non-Treasury deposits Domestic bank deposits at Federal Reserve banks (FS, 1963, pp. 737–740) are used.

Net foreign assets (other than foreign-held currency) Holdings of foreign bills *plus* amount due from foreign banks *minus* foreign deposits at Federal Reserve banks (Governors, 1943, pp. 330–332) are used.

Lost and foreign-held currency Consider the FS monetary base series for end-of-November 1874–1906, end-of-February 1874–1907, and end-of-year

1907–1932. For comparability with the new monetary base, state bank notes are excluded. The FS source for state bank notes is Governors (1943, p. 408), which shows $1 million in circulation June 1873 to June 1878 and then zero. So $1 million is deducted from the November figures 1874–1877 and February figures 1874–1878. Linear interpolation, as in note 24, is applied to obtain an end-of-year series 1874–1932 (hereafter 'the adjusted FS series'). Estimated Treasury and Federal Reserve currency in official circulation 1914–1932 is obtained by subtracting (1) specie in circulation (specie stock [from section A] *minus* Federal Reserve gold *minus* Treasury gross specie [from section G]), (2) nongold coin in circulation (from section G), and (3) domestic bank deposits at Federal Reserve banks. For 1914–1932, one computes the annual ratio of Federal Reserve currency to the estimated sum of Treasury and Federal Reserve currency in official circulation. This ratio multiplies 'lost currency net of lost currency in 1913' (section C) to yield lost Federal Reserve currency. It multiplies 'foreign-held currency net of foreign-held currency in 1913' to estimate foreign-held Federal Reserve currency.

G. Treasury variables

Specie The position that Treasury specie is zero for 1789–1835 is irrefutable (Treasury *Report*, 1915, p. 374; FS, 1970, pp. 245–246). It is the sense of FS (1970, p. 248) that this is true also for 1836–1846, which appears to be the position of Taus (1943, pp. 222–224), except for the period of the first Independent Treasury. Therefore, through 1846, Treasury specie is taken as zero except for end-of-year 1840, where gold is assumed to constitute half of the balances in Treasury offices (from Treasury *Report*, 1915, p. 374). For end-of-year 1847–1859, the Macesich data published in FS (1970, pp. 222–224) are used.

For 1860–1861 and 1864–1872, midyear data are computed as the (uncorrected) stock of gold coin and bullion *minus* circulation of gold coin from Treasury *Report* (1928, pp. 552–554). The ratio of Treasury gold to the corrected gold stock for midyear 1861 and 1864 is interpolated linearly to obtain midyear ratios for 1862–1863, which in turn multiply the corrected stock to estimate midyear Treasury gold for these years. For 1873–1878, midyear Treasury gold is from Mint *Report* (1941, p. 84). Midyear averaging yields Treasury gold end-of-year 1860–1877. The figure for end-of-year 1878 is in Treasury *Report* (1898, p. 59). Treasury gold for end-of-year 1879–1932 is from Mint *Report* (1941, p. 84).

Gold certificates Gold certificates were first issued in 1865, but in a trivial amount (Bayley, 1881, p. 162). For 1866–1877, midyear averaging is performed on official circulation data in Treasury *Report* (1928, p. 554). End-of-year data 1878–1932 are from Governors (1943, pp. 409–412) and Treasury *Report* (1898, pp. 131–132; 1903, pp. 219–220; 1909, pp. 204–208; 1915, pp. 351–354).

Nongold coin in circulation The sum of silver dollars, subsidiary silver coin, and minor coin is used. Silver dollars were not in circulation during 1860–1877. However, following FS (1963, pp. 113–114, n. 37; 723), the 1877 figure is taken as $6 million, representing circulation of trade dollars. Standard silver dollars in circulation are available end-of-year 1878–1932 (Governors, 1943, pp. 409–412; Treasury *Report*, 1898, pp. 124–126; 1903, pp. 215–216; 1909, pp. 194–198; 1915, pp. 343–346).

Considering subsidiary silver coin, for 1860–1873 midyear averaging is applied to data from the ultimate source: Treasury *Report* (1928, pp. 552, 554). For midyear 1860–1863, only the stock figure is available, and the 1864 circulation/stock figure is used to estimate circulation. For 1874–1877, midyear averaging is applied to figures in Governors (1943, p. 408), the FS data source (containing fewer significant digits than Treasury *Report*, 1928). For 1878–1932, the sources are the same as for silver dollars. Following FS (1963, p. 723; 1970, p. 355), the overestimate deducted in mid-1910 by the Director of the Mint is apportioned linearly over 1881–1909.

Minor coin in circulation is available midyear 1900–1914 and end-of-year 1914–1932 (Governors, 1943, pp. 408–412). Midyear averaging is applied to the former.

Currency, 1812–1873 For Treasury notes (1812–1873), outstanding issues are taken from public debt statements: end-of-year 1812–1843 from Elliot (1845, pp. 906–917) and various end-of-quarter dates 1844–1874 from Treasury *Reports*. The latter figures are converted to end of year via (1) addition of quarterly issues *minus* redemptions from Bayley (1881) or (2) linear interpolation, used where the two adjacent known figures both are below $1 million (in fact, below $0.65 million) *and* issues are zero during the interpolation period.[49] For old demand notes, Treasury notes of 1863, compound interest notes, and 3% certificates (1861–1873), end-of-year figures for the initial year are the issues (with redemptions again zero) (Bayley, 1881, pp. 153, 161–163, 169). Then midyear-averaged figures in outstanding public debt statements (Treasury *Reports*, 1862–1874) are taken.[50] For U.S. notes (greenbacks, 1862–1873), official circulation is obtained via midyear averaging of data in Treasury *Report* (1928, p. 554). For fractional currency (1862–1873), the same applies, with two amendments. First, following FS (1963, p. 724; 1970, pp. 354–355), all but $1 million of outstanding fractional currency in midyear 1878 is assumed lost, distributed linearly over 1863–1878, and deducted from the official data. Second, the initial (end-of-1862) figure is total issues during 1862 (there were no redemptions [Bayley, 1881, pp. 159–160]) multiplied by the circulation/stock ratio of midyear 1863, with the estimated loss subtracted.

The sum of all the above components *plus* gold certificates in official circulation (from 1866) *minus* lost currency (section C) yields Treasury currency (with no gold certificates) in domestic circulation 1812–1865 but inclusive of gold certificates 1866–1873.

Currency, 1874–1932 The adjusted FS series *minus* specie in circulation (section F) *minus* lost currency and foreign-held currency (section C) *minus* nongold coin in circulation yields Treasury currency, inclusive of gold certificates, in domestic circulation 1874–1913. The series for 1914–1932 is obtained as this result *minus* Federal Reserve currency in official circulation *plus* lost Federal Reserve currency *plus* foreign-held Federal Reserve currency *minus* domestic bank deposits at Federal Reserve Banks (section F).

Currency, exclusive of gold certificates, in domestic circulation, 1866–1932 This is obtained by *subtracting* gold certificates in official circulation and *adding* lost gold certificates. Lost certificates are the product of (1) the ratio of official circulation of gold certificates to that of old demand notes, U.S. notes, national bank notes (from 1874), silver certificates, Treasury notes of 1890, and gold certificates, and (2) lost Treasury currency inclusive of gold certificates, net of lost currency in 1865.

For 1866–1873, the denominator of the ratio consists of old demand notes, U.S. notes, and gold certificates. For 1874–1913, the denominator is estimated as the adjusted FS monetary base *minus* specie in circulation *minus* nongold coin in circulation. For 1914–1932, Federal Reserve currency in official circulation and domestic bank deposits at Federal Reserve Banks (section F) are also subtracted. Lost Treasury currency is total lost currency (section C) *minus* lost Federal Reserve currency (section F).

H. Prices

Specie price of currency For the central bank, this variable is unity except for May 1837 to August 1838, when the Second Bank suspended specie payments. The percentage premium (PR) on American gold at Philadelphia for end-of-year 1837 is linearly interpolated between December 9, 1837, and January 6, 1838, observations (SD457 25C 2s). The specie price of currency is then 1/(1 + PR/100) = .9609. Nonunity specie price of currency for the Treasury is the gold price of greenbacks for the last market day of the year, 1861–1878.[51]

Price level The paper price level is measured by the GNP deflator. For 1792–1869, the source is Berry (1988, p. 21), ratio linked in 1869 to the series for 1869–1932 in Balke and Gordon (1989, pp. 84–85) and Department of Commerce (1986, pp. 1, 6). The gold price level (P) is the product of the paper price level and the specie price of currency (for the full year rather than end of year), with par equaling unity.

The annual specie price of currency for the antebellum period is derived as follows. The Berry deflator is based on the Hoover and Taylor (1959) composite index of wholesale price indexes in various cities. Let $PCUR^i$ denote the specie price of currency in city i. The weighting pattern of the Hoover-Taylor index (differing for 1800–1815 from 1816–1861) is applied to the data-available $PCUR^i$ for periods during which at least one city is on a paper standard ($PCUR^i < 1$), based on information in Officer (1996, pp. 16–17) and Berry (1943). Thus, the specie price of currency is a weighted average of $PCUR^i$ for New York and Philadelphia, (1814–1817); Cincinnati and the other cities (for which $PCUR^i = 1$) (1818–1820); New York, Philadelphia, and Cincinnati (1837–1842–but $PCUR^i = 1$ for New York 1839–1842); and New York, Cincinnati, and New Orleans (for which $PCUR^i = 1$) (1857).[52] For 1862–1878, the specie price of currency is the gold price of greenbacks (Mitchell, 1908, p. 4).

I. Income

The logical source for income (Y) 1792–1869 is the real GNP series of Berry (1988, pp. 18–20), consistent with construction of the price level. However, Berry's technique is subject to legitimate criticism for the antebellum period.[53] Fortunately, the limitations of Berry's series are overcome via the broad concept real GDP data of Weiss (1992, pp. 31–32). The Weiss figures, developed for nine antebellum benchmark years (1793, 1800, 1807, 1810, 1820, . . ., 1860), are on a per capita basis. Multiplication by population (Census, 1975, p. 8) yields YW, the Weiss gross domestic product (GDP) series.

Denoting the Berry series as YB, a revised series (YR) is derived as follows: (1) For 1860–1869, YR = YB. (2) Running t from 1850 back in time over the benchmark years (with successive such years separated by m calendar years),

$$YR_t = \frac{YW_t}{YW_{t+m}} \cdot YR_{t+m}. \quad (3)\ YR_{1792} = \frac{YR_{1793}}{YB_{1793}} \cdot YB_{1792}.$$

$$(4)\ \text{Let } f = \frac{\left(\dfrac{YR_{t+m}}{YB_{t+m}} - \dfrac{YR_t}{YB_t}\right)}{m}.$$

Then $YR_{t+n} = ((YR_t/YB_t) + n \cdot f) \cdot YB_{t+n}$, $t = 1850, 1840, \ldots, 1 \le n < m$. The source of income 1869–1932 is the same as for the paper price level. Balke and Gordon (1989) take care to express real GNP consistent with the national accounts (Department of Commerce, 1986) denomination in 1982 constant dollars, whence the price level equals 100 for that year.[54] The revised series for 1792–1869 is ratio linked to the Balke-Gordon series in 1869. Per capita income is the ratio of real income to population.

J. Foreign variables

The foreign variables are index numbers: Britain (converted to 1913 = 1) 1791–1913, an index of Britain (.5778 weight) and Canada (.4222 weight) (both converted to 1913 = 1) 1913–1932. Weights are proportional to share of U.S. exports and imports during 1913–1932 (Census, 1975, pp. 903–906).

Exchange rate (E) For 1791–1913, the exchange rate is based on the annual average of the quarterly sight bill equivalent exchange rate (dollars per pound) corrected for paper currency depreciation (obtained by reversing the procedures in Officer, 1996, pp. 54–55, 64–97). The inverse of this series (whence pounds per dollar) is taken and expressed as 1913 = 1.

For 1913–1932, annual cable exchange rate (DP = dollars per pound and DC = dollars per Canadian dollar) are from Governors (1943, pp. 665, 681). The United Kingdom was on a paper standard from August 1914 to April 27, 1925, and again from September 20, 1931, as was Canada to June 30, 1926, and from January 1929. It may be noted that dollar/sterling and Canadian-dollar/sterling parity was 4.8665635, with Canadian-dollar/U.S.-dollar parity at unity. The London gold market was closed during the paper standard until September 1919. So the exchange rates are corrected for paper currency depreciation 1919–1925 and 1931–1932 as follows.

Letting PGL denote the currency price of gold in London (the ratio of the market price of gold [from Shrigley, 1935, p. 92] to the mint parity price of gold), PRP = (1/PGL − 1) is the proportionate premium of the pound over gold (with the pound at a discount, PRP is negative). The corrected dollar/pound exchange rate is DPC = DP – 4.8665635 · PRP. Letting CP denote the Canadian-dollar/pound cable exchange rate (from Leacy, 1983, series J563), the proportionate premium of the Canadian dollar with respect to gold is PRC = (4.8665635/CP) · (1/PGL) − 1, and the corrected dollar/Canadian-dollar exchange rate is DCC = DC – PRC. The inverses of DPC and DCC are then expressed in index number form.

Price level (P^f) Considering Britain for 1790–1830, the Gayer, Rostow, and Schwartz price index (in Mitchell, 1988, p. 721) is ratio-linked to the GDP deflator for 1830–1932. The latter is constructed as the ratio of current price to constant price GDP, with the numerator and denominator each obtained by ratio-linking earlier to the first year of later component series: Feinstein (in Mitchell, 1988, pp. 831–838) expenditure (1830–1854) and "compromise" (1855–1869, 1913–1919) estimates of GDP at factor cost, Solomou and Weale (1991, p. 60; 1996, pp. 110–113) 'balanced' estimate of GDP (1870–1912 and 1920–1932). The Canadian gross national product (GNP) deflator is from Urquhart (1993, p. 25) 1913–1926, ratio-linked to Statistics Canada (Leacy, 1983, series K172) 1926–1932.

For each country, the gold price level is the product of the paper price level and the gold price of currency. The currency price of gold in London (PGL) must be extended to 1797–1821, the Bank Restriction Period of the paper pound. Quarterly averages of the price of bar gold are computed from weekly observations in *Report* ([1819] 1968, pp. 335–354) for 1797–1818 and *Report* ([1832] 1968, pp. 98–100) for 1819–1821. Annual averages of the available quarterly observations are taken, and linear interpolation is applied for missing years (1800–1803 and 1806–1809).[55] The gold price of currency for Britain is 1/PGL, while for Canada it is (PRC + 1).

Monetary base (*BASE^f*) The Canadian monetary base (1913–1932), from Metcalf, Redish, and Shearer (1996), is conceptually equivalent to the FS base and the new monetary base. The British series ($BASE^B$), developed here, differs in including only domestic bank deposits (hereafter 'bankers' balances') at the Bank of England, excluding other non-central government deposits because the latter cannot be separated from foreign deposits.

For 1791–1869, $BASE^B$ is the sum of coin in circulation (CC), Bank of England notes in circulation (BN), Scottish and Irish banks notes in circulation *less* coin held (SIN) (from 1845, pursuant to the Bankers' Acts [Scotland and Ireland] of that year), and bankers' balances (BB). The sources for SIN are *Report* ([1857] 1969) for (last date in year) 1845–1856 and *The Economist* (4-week average ending date closest to year-end) 1857–1869. For 1791–1867, CC = SP – BAC, where SP is the specie stock and BAC is the Bank of England coin and bullion. CC 1868–1869 is midyear-averaged figures of Capie and Webber (1985, p. 198). For 1844–1869, BAC and BN (constructed as notes issued *minus* notes in Banking Department) are from *The Economist*, closest return to end of year. Prior to 1844, the preferred source of any Bank series is Bank of England *Quarterly Bulletin* (June 1967, Appendix [hereafter QB]). Other series for BAC and BN are in *Reports* ([1840, 1841, 1848] 1968).

Let QBF denote the QB end-of-February series, (RF, RN, RD) the corresponding *Report* series for end of (February, November, December), and the subscript 1 the series forwarded one year. Formula A is $(QBF_1/RF_1) \cdot RD$, formula B differs in linearly interpolating RD as $(2{\cdot}RN + RF_1)/3$, and formula C is $(2{\cdot}QBF + 10{\cdot}QBF_1)/12$. Formula A is used to estimate BN 1792–1797, BN 1815–1843, and BAC 1832–1843; formula B to estimate BN 1791, BN 1798–1814, and BAC 1816–1831; formula C to estimate BAC 1791–1815.

Benchmark year-end dates for SP are 1790 (the 1800 figure *minus* 10 times average annual net imports of specie 1791–1800 [from Brezis, 1995, p. 51]); 1800, 1830, and 1860 (from Feinstein, 1988, p. 397); and 1868 (composed as CC + BAC). Net imports of specie (F^B) are from Brezis 1791–1800, computed as $(SP_{1830} - SP_{1800} - \sum_{1816}^{1830} F^B)/15$ for 1801–1815, and from Imlah (1958, pp. 70–72), changing sign of his net exports series, for 1816–1868. The interpolative technique for SPST is then applied to SP, with F^B the interpolative series.[56]

For 1791–1818, BB is constructed as U · V · TD, where TD is total deposits, *V* is the estimated ratio of private (non-central government) deposits to total deposits, and *U* is the estimated ratio of bankers' balances to private deposits. Data are from QB and *Report* ([1832] 1968). TD is obtained by applying formula (C and B) to (1790–1814 and 1815–1818). Considering the numerator and denominator of *V*: for 1791–1806, they are the annual average of 1807; for 1807–1813, they are the annual average of the current year plus the annual average of the subsequent year; for 1814, they pertain to February

1815; for 1815–1818, they are linearly interpolated as for RD in formula B. For 1791–1818, $U = BB_{1819}/(V_{1819} \cdot TD_{1819})$.

For 1819–1869, BB is estimated as $(BB_{1870}/BBH_{1870}) \cdot BBH$, where BBH is bankers' balances at the Bank head office (from QB) and BB (from Capie and Webber, 1985, p. 409) also includes balances at branches.[57]

For 1870–1932, $BASE^B$ is obtained from the Capie and Webber (1985, pp. 54–57) end-of-year series by *adding* Bank of England Banking Department coin (last reporting date in December, from *The Economist*) and *subtracting* Banking Department notes and coin (Capie and Webber, 1985, pp. 409–420).[58]

Notes

* Reprinted from *Explorations in Economic History*, vol. 39, Lawrence H. Officer, 'The U.S. Specie Standard, 1792–1932: Some Monetarist Arithmetic,' pp. 113–153. Copyright 2002, with permission from Elsevier.

1 The author thanks Michael Bordo and Forrest Capie for assistance in obtaining Bank of England data.

2 Rutner (1974), who has performed the most thorough investigation of the central bank status of the Second Bank, states, 'The ultimate criterion by which the BUS could be a central bank and which would make it unique is simply this: did other economic actors (i.e., banks and individuals) consider BUS monetary liabilities a form of reserve currency?' (p. 121). He answers strongly in the affirmative (see below).

3 See Hammond (1957, p. 403), Warburton (1962, p. 67), Fenstermaker (1965, p. 69), Rockoff (1971, p. 456), Rutner (1974, pp. 23, 27, 143–144), Timberlake (1993, p. 241), and Highfield, O'Hara, and Smith (1996, p. 483). However, Temin treats the Second Bank as a commercial bank, albeit an important one.

4 The fact that they were not uniformly so included is not a 'puzzle' (Rutner's term), for (1) Bank notes were not a legal reserve and (2) there was no minimum reserve requirement.

5 By contrast, Fenstermaker and Filer (1986) find that the Banks of the United States did not affect the behavior of New England state banks, but they view this result as purely regional.

6 Rutner (1974) observes that even 'in the Panic of 1837 . . ., there appears to be fairly strong evidence to suggest that the BUS monetary obligations were considered a form of reserve currency and hence in this sense the BUS was a central bank' (p. 145).

7 Within a few months of beginning operations at Philadelphia (the head office), each Bank established branches in Baltimore, Boston, Charleston, and New York (plus 13 other locations on the part of the Second Bank). Ultimately, the First Bank had eight branches, and the Second Bank had a maximum of 26 at one time. After the Second Bank became a Pennsylvania state bank, it continued to operate nationally by converting its branches to agencies.

8 The monetary control argument is best made for the First Bank by Hammond (1957, pp. 198–199) and Perkins (1994, p. 249) and for the Second Bank by Temin (1969, pp. 49–53) and Timberlake (1993, p. 241). Rockoff (1971, pp. 456–457) observes that the Second Bank continued this form of monetary regulation even after it became a Pennsylvania state bank.

9 This comparison, made by historians for the Second Bank (e.g., Shultz and Caine, 1937, p. 211; Smith, 1953, p. 236; Studenski and Krooss, 1963, p. 87), again can be extended to the First Bank.

10 The Bordo–Kydland–Rockoff thesis suggests that it is a mistake to view the greenback period (or any suspension of specie payments) as uniformly involving the weakest adherence to a metallic standard and to view the classic gold standard as uniformly involving the strongest. In fact, for most of the last decade of the greenback period, there was strong expectation of a return to the former gold standard (Bordo and Kydland, 1995, pp. 451–452), and for much of the early and mid-1890s, there was a high objective and subjective probability of U.S. abandonment of the gold standard (FS, 1963, pp. 104–113).

11 It is arguable that the First and Second Banks gained their central bank status only gradually when the institution came into existence and lost it similarly when the Bank was on its way out. In this vein, Rutner (1974) asks, 'Did BUS [Second Bank] monetary liabilities lose their "high-poweredness" in a continuous or discontinuous manner?' (p. 125). He includes Second Bank notes and deposits in the monetary base until the very end of the Bank's existence in early 1842, but he also shows an alternative series excluding the Bank's liabilities from the base. The ideal solution might be to assign weights to the Banks' liabilities increasing from zero to unity at the beginning, decreasing from unity to zero at the end, if only the weighting patterns were known. The current study, in effect, allocates a weight of unity to Second Bank liabilities until the end of 1838 and a zero weight from the end of 1839.

12 Because the First and Second Banks did not generally behave as lenders of last resort, they were 'outside' agents only in the sense of having their liabilities serve as components of the monetary base. However, it is also true that the performance of the Federal Reserve System as a lender of last resort during the early 1930s was 'little more than lip service' (FS, 1963, p. 395).

13 Prior to 1860, nongold coin in circulation is included in the specie stock.

14 Currency is at par when measurable amounts are held by foreigners. So there is no conversion process for foreign-held currency.

15 Throughout this study, beginning-of-year data are considered end-of-previous-year data.

16 Yet there remains a conceptual problem. The greenback price of gold is highly correlated with the price level, and depreciated monies constitute 69% of the base during the greenback period (see Tables 11.2 and 11.3). So the gold-denominated base is roughly the real base for this period. Certainly, one would not apply this procedure after 1932, and especially after 1972, when the paper dollar nominal monetary base (constituting the entire base) would be deflated by a volatile price of gold. So legitimate comparisons between the *greenback period* new monetary base and the post-1932 base could not be readily made. Also, in the long run, the resulting new base might approximate the real base, and a nominal base does not remain for analysis. I am indebted to an anonymous referee for raising several important issues, including this one.

17 In principle, as a compromise between the two positions, national bank notes could enter the monetary base prior to 1874 but with a weight below unity.

18 Notwithstanding the Mint Act of June 28, 1834, which undervalued silver relative to gold, there is evidence that 'silver coins remained in common use in the United States until some time after the discovery of gold in California [in 1848]' (Berry, 1943, p. 488). In a similar vein, Martin (1973) shows that 'de facto bimetallism . . . persisted to mid-century' (p. 825). It appears that the turning point was the Subsidiary Coinage Act of February 21, 1853 (Officer, 1996, p. 20), but Berry observes that as late as 1857, silver (along with gold) coin was advertised at a premium.

19 Smith (1953) refers to 'the post notes of ill repute' and observes that 'the amount of these issues was a reliable index of the degree of financial emergency within the Bank' (p. 182).

20 The FS (1963, p. 25, n. 10) claim that their currency figures include 3% certificates is false; see note 50. However, FS are followed in their *exclusion* from the monetary base of other interest-bearing currency issued during the Civil War. Recent assessments of the 'moneyness' of various forms of interest-bearing Civil War currency are in Gherity (1993) and Woodward (1995). The evidence is mixed and intertwined with the definition of moneyness.

21 Carothers (1930, pp. 170–185, 241–261) provides the best history of these remarkable currencies. He observes that 'these glue-coated bits of paper [postage currency] were the worst form of currency ever used by a civilized people' (p. 174) and, quoting Knox, that fractional currency 'wore out rapidly and became ragged and filthy' (p. 184).

22 See FS (1963, pp. 207, 257, n. 40), Rutner (1974, pp. 248–253), and Sylla (1982, pp. 31–33).

23 Also, none of the authors includes the Civil War years, and only Bordo includes the antebellum period. Furthermore, FS provide charts rather than figures; Bordo and Cagan deal only with *changes* in the base; and, like Temin, Bordo defines the antebellum monetary base as composed only of specie.

24 Temin's (1969, pp. 186–187) series is at end of fiscal year (September 30, 1820–1842, and June 30, 1843–1858). Rutner's series (not seasonally adjusted, with Second Bank a central bank) is selected for compatibility with the new base. It has year-end data points except for 1835, 1840, and 1843–1846. FS (1963, pp. 800–804; 1970, pp. 344–350) provide data for end-of-November 1867–1906, end-of-February 1867–1907, and end-of-year 1907–1932; the November–February figures serve as interpolative points for year-end figures 1867–1906.

25 The formula to calculate annual average percentage change in X is $100 \cdot \ln(X_{t+n}/X_t)/n$, where t is the initial year and $t + n$ is the final year.

26 Taking the first difference of equation (1) and incorporating equations (2) and (3) yields $\Delta BASE_S = BP + \Delta DOB$.

27 Throughout the model, price levels and the exchange rate are corrected for paper currency depreciation, in conformity with the monetary base expressed in gold dollars.

28 The exchange market pressure model, of which equation (6) is a generalization, has been criticized by Weymark (1995, 1997a, 1997b, 1998). She argues that EMP is the simple sum of the change in official reserves (here, balance of payments) and exchange rate components only under restrictive assumptions: (1) purely monetary model, (2) small open economy, (3) exogenous exchange rate intervention, and (4) exogenous change in domestic credit (a component of ΔDOB in the current model). In response, first, Weymark extends the monetary model by incorporating aggregate demand and supply (and nontraded goods), but at the cost of complexity. The monetary approach is readily operational and melds well with the monetary balance of payments. Second, introducing foreign country exchange market intervention into EMP is unsuitable for the purpose at hand, in which EMP is constructed to impinge directly on the domestic economy. Therefore, the totality of the foreign base term may be placed in SB. Third, under a specie standard, exchange rates are kept within the gold point spread typically by passive specie transactions behavior of the authorities. Fourth, providing that the authorities respond to EMP itself or that they sterilize gold flows, the definition of EMP remains valid even with endogenous change in domestic credit. In fact, the Second Bank under Biddle altered domestic credit in response to both specie flow and exchange rate change (Redlich, 1968, pp. 125, 134), and the Federal Reserve System sterilized gold flows for much of the 1920s and into the second half of 1931 (FS, 1963, pp. 279–287, 297, 360–361, 396–399).

29 In the construction of variables, the proportionate change in X is $\Delta X/X_{-1}$.

30 It is also true that 1879–1913 has the smallest magnitude of the absolute value of

every component of EMP, no matter how composed: BP^{dir}/BASE, BP^{ind}/BASE, $\Delta E/E$, DPP, DB, SB^{dir}, and SB^{ind}.

31 This statement is valid only if P and Y are defined so that (1) $P = 1$ in the national accounts base year and (2) the unit of measurement of Y is the same as that for BASE. Otherwise, the ratio is income velocity only up to a multiplicative constant. P and Y are constructed to make the statement true.

32 Nearly all of the studies are listed in Bordo and Schwartz (1999). See also Basu and Taylor (1999).

33 This is the technique of Temin (1969, pp. 185–189) and Rutner (1974, pp. 205–216) as well as that of Seaman (1852, pp. 257–260); Secretary of the Treasury, *Annual Report* [hereafter 'Treasury *Report*'] (1855, p. 71); and Warburton (reported in FS, 1970, p. 227).

34 See Treasury *Report* (1854, p. 281; 1855, p. 71) and Simon (1960, pp. 631–632, 644).

35 Blodget's series, for 1790–1807, is dated end of year by FS (1970, pp. 216–219), but it is interpreted as beginning of year (end of the previous year) by Temin (1969, p. 185), and by FS (1970, p. 244, n. 16) themselves, via the dating of the table in Treasury Department (1915, p. 45), which includes the Blodget figure for 1800.

36 Rutner (1974, pp. 205–207) believes that this is Woodbury's basic figure, and FS (1970, p. 227) provide evidence that it is indeed so, but it is uncertain whether the estimate is derived purely from stock data. Woodbury describes his numbers only as 'prepared partly from actual returns, and partly from estimates' (Elliot, 1845, pp. 941–942).

37 The source is Census (1949, pp. 243–245), with 'calendar-year annualization' of figures for other than calendar years. For example, data for years ending June 30 (September 30) are allocated 50% (75%) to the current (the remainder to the previous) year.

38 Linear interpolation is used between benchmark dates, and a half-year of operation is assumed for the initial year (1836) of the only significant silver mine. Data are converted from physical output to value via multiplication by the New York price of silver (1836–1849 from Director of the Mint, *Annual Report* [hereafter 'Mint *Report*'], 1910, p. 99, with price computed as the ratio of value to output; 1850–1860 from Census, 1975, p. 606). Rutner uses Herfindahl's silver (and gold) data, but only for 1834–1849.

39 Calendar-year annualization is applied as warranted. Temin makes no allowance for nonmonetary consumption. Rutner and Shetler (1973) do not employ pre-1880 consumption data for their antebellum estimates (and Rutner errs in including reworked metal), but Seaman shows an appreciation of the concept of nonmonetary consumption that vindicates his numbers. Seaman's figures for 1821–1846 are net of domestic production and require restoration to gross level. From the text, it may be inferred that he takes production as (1) essentially zero for 1821–1823, (2) deposits of domestic gold production at the mints for 1824–1829, and (3) \$500,000 for 1830–1846. The figure for 1820 is obtained by assuming that the percentage decline in consumption from 1821 to 1811 was the same as that from 1831 to 1821 and applying linear interpolation.

40 This is an important antebellum interpolation. It is prudent to check whether log–linear rather than linear interpolation makes a difference. The Theil inequality coefficient between the alternative interpolative series and the actual series F is 0.0086, with zero being a perfect fit.

41 Note that linear interpolation involves (j/n) in place of the bracketed term.

42 The source is Laurent (1974, p. 221). It is reasonable to assume that large-denomination notes would be guarded most carefully.

43 Garber (1986, pp. S140–S141, S150) provides evidence that 'prior to World War I little U.S. currency was held in Europe.' It is unlikely that this situation

changed until some time after American Expeditionary Forces arrived in France in June 1917.

44 This is quantitatively the most important linear interpolation for foreign-held dollars. Log–linearity would change the monetary base in any year by less than \$6 million, less than one-fourth of 1%.

45 A Honduras figure for 1922 is unreasonably low and so is disregarded.

46 This is the view of Perkins (1994, p. 248), who computes the ratio for various years.

47 The exchange rate to convert foreign bills in 1793 from guilders to dollars is in Wettereau (1985, p. 87).

48 The sources are HD52 17C 1s (1821), HD78 18C 1s (1822–1823), HD105 19C 1s (1825), American State Papers Finance 766 19C 2s (1826), HD100 20C 1s (1827), HD93 20C 2s (1828), HD63 21C 2s (1829–1830), HD523 23C 1s (1831–1833), SD128 25C 2s (actual returns) (1834, 1836), SD312 24C 1s (1835), SD 471 25C 2s (condensed return) (1837), and HD172 26C 1s (1838). Smith did not locate returns prior to 1825.

49 There are discrepancies between Bayley's flow data and the change in amount outstanding obtained by first-differencing the public debt series, but the divergence is of importance only for small changes in amount outstanding. For possible reasons for the discrepancy, see Treasury *Report* (1846, p. 29) and Rutner (1974, p. 253). Bayley's figures are probably superior to the Treasury flow data – the latter used by Rutner – because Bayley accounts for and corrects anomalies in the Treasury data, Also, Rutner obtains his outstanding notes series by continuously cumulating sales *minus* retirements, a technique that fails to take advantage of the (presumed definitive) public debt statements.

50 In using the 'other U.S. currency' series rather than consulting the public debt statements, FS commit actual or potential errors. First, 'other U.S. currency,' as found in Treasury *Report* (1928, p. 552), equals the sum of outstanding old demand notes, Treasury notes of 1863, and compound interest notes; 3% certificates are excluded. Second, for midyear 1863, 'other U.S. currency' is overstated by including (and thus double-counting) the stock of U.S. notes issued under the Act of March 3, 1863. At \$89.879 million, the error is substantial – 20% of the 1862–1863 average monetary base. By 1874, 3% certificates outstanding are nearly zero, and there is only a trivial difference between 'other U.S. currency' and the sum of the components in the public debt statement.

51 The average of the high and low price for the day is taken, from Mitchell (1908, pp. 288–338). For 1861, the January 1, 1862, figure is used.

52 Sources of $PCUR^i$ are Gallatin (1831, p. 106) for 1814–1817, Warren and Pearson (1935, p. 154) for New York 1837–1838 and 1857, Officer (1996, p. 78) for Philadelphia 1837–1842, and Berry (1943, pp. 386–389, 398, 462, 590–591) for Cincinnati. Averages of monthly or quarterly values, often of the percentage specie premium, are taken. (Where there is a monthly range, the midpoint is used.) If the annual value of the specie premium is PR^i, then $PCUR^i = 1/(1 + PR^i)$. Berry provides no data for June 1839 to March 1840, but the specie premium is clearly zero for June-September. He notes that the specie premium increased to about 8% within 5 months of the October 1839 suspension. It is assumed that the premium increased linearly from zero just prior to that suspension to 8% in March 1840, and a suitable weighted average of zero and the interpolated value is computed for October 1839.

53 As noted by Engerman and Gallman (1982, pp. 5, 15–16), the extrapolator series are few in number, the GNP concept excludes home production, and the extrapolations are based on a statistical model devoid of economic content.

54 Balke and Gordon (1989, p. 40) argue convincingly for their own superiority over the competing Romer (1989, pp. 22–23) series. Dividing P by 100 and expressing Y in millions rather than billions of dollars satisfies (1) and (2) in note 31.

55 This technique results in a series superior to those of Tooke (in Arnon, 1991, p. 159) and Hawtrey (1918, p. 64).

56 For 1791–1800 and 1801–1815, by construction, $\sum_{i=1}^{N} F_i^{B} = \Delta_0^n$ SP.

57 For 1819–1827, QB data are beginning of subsequent year.

58 Thus, the Capie-Webber series is corrected for, inconsistently, excluding Banking Department coin but including its notes.

References

Arnon, A. (1991), *Thomas Tooke*. Ann Arbor: Univ. of Michigan Press.

Balke, N. S., and Gordon, R. J. (1989), 'The Estimation of Prewar Gross National Product: Methodology and New Evidence.' *Journal of Political Economy* **97**, 38–92.

Bank of England, *Quarterly Bulletin* (June 1967, Appendix).

Basu, S., and Taylor, A. M. (1999), 'Business Cycles in International Historical Perspective.' *Journal of Economic Perspectives* **13**, 45–68.

Bayley, R. A. (1881), *National Loans of the United States*. Washington, DC: Government Printing Office.

Berry, T. S. (1943), *Western Prices before 1861*. Cambridge, MA: Harvard Univ. Press.

Berry, T. S. (1984), *Early California Gold, Prices, Trade*. Richmond, VA: Bostwick.

Berry, T. S. (1988), *Production and Population since 1789*. Richmond, VA: Bostwick.

Board of Governors of the Federal Reserve System (1943), *Banking and Monetary Statistics, 1914–1941*. Washington, DC.

Bordo, M. D. (1975), 'The Income Effects of the Sources of Monetary Change: A Historical Approach.' *Economic Inquiry* **13**, 505–525.

Bordo, M. D., and Kydland, F. E. (1995), 'The Gold Standard as a Rule: An Essay in Exploration.' *Explorations in Economic History* **32**, 423–464.

Bordo, M. D., and Rockoff, H. (1996), 'The Gold Standard as a "Good Housekeeping Seal of Approval," ' *Journal of Economic History* **56**, 389–428.

Bordo, M. D., and Schwartz, A. J. (1999), 'Monetary Policy Regimes and Economic Performance: The Historical Record.' In J. B. Taylor and M. Woodford (Eds.), *Handbook of Macroeconomics*, vol. 1A. Amsterdam: Elsevier.

Brezis, E. S. (1995), 'Foreign Capital Flows in the Century of Britain's Industrial Revolution: New Estimates, Controlled Conjectures.' *Economic History Review* **48**, 46–67.

Bureau of the Census (1949), *Historical Statistics of the United States, 1789–1945*. Washington, DC: Government Printing Office.

Bureau of the Census (1975), *Historical Statistics of the United States, Colonial Times to 1970*. Washington, DC: Government Printing Office.

Cagan, P. (1965), *Determinants and Effects of Changes in the Stock of Money, 1875–1960*. New York: Columbia Univ. Press.

Capie, F., and Webber, A. (1985), *A Monetary History of he United Kingdom, 1870–1982*. London: Allen & Unwin.

Carothers, N. (1930), *Fractional Money*. New York: Wiley.

Department of Commerce (1986), *The National Income and Product Accounts of the United States, 1929–82: Statistical Tables*. Washington, DC: Government Printing Office.

Director of the Mint (various years), *Annual Report*.

Elliot, J. (1845), *The Funding System of the United States and of Great Britain*. Washington, DC: Blair & Rives.

Engerman, S. L. (1970), 'A Note on the Economic Consequences of the Second Bank of the United States.' *Journal of Political Economy* **73**, 725–728.

Engerman, S. L., and Gallman, R. E. (1982), 'U.S. Economic Growth, 1783–1860.' in P. Uselding (Ed.), *Research in Economic History*, Vol. 8. Greenwich, CT: JAI.

Feinstein, C. H. (1988), 'National Statistics, 1760–1920.' In C. H. Feinstein and S. Pollard (Eds.), *Studies in Capital Formation in the United Kingdom, 1750–1920*. Oxford, UK: Clarendon.

Fenstermaker, J. V. (1965), *The Development of American Commercial Banking: 1782–1837*. Kent, OH: Kent State Univ., Bureau of Economic and Business Research.

Fenstermaker, J. V., and Filer, J. E. (1986), 'Impact of the First and Second Banks of the United States and the Suffolk System on New England Bank Money: 1791–1837.' *Journal of Money, Credit and Banking* **18**, 28–40.

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

Friedman, M., and Schwartz, A. J. (1970), *Monetary Statistics of the United States*. New York: National Bureau of Economic Research.

Gallatin, A. (1831), *Considerations on the Currency and Banking System of the United States*. Philadelphia: Carey & Lea.

Garber, P. M. (1986), 'Interwar Movements of Dollars to Europe and the U.S. Currency Supply.' *Journal of International Money and Finance* **5**, S135–S156.

Gherity, J. A. (1993), 'Interest-Bearing Currency: Evidence from the Civil War Experience.' *Journal of Money, Credit and Banking* **25**, 125–131.

Hammond, B. (1947), 'Jackson, Biddle, and the Bank of the United States.' *Journal of Economic History*, **7**, 1–23.

Hammond, B. (1957), *Banks and Politics in America*. Princeton, NJ: Princeton Univ. Press.

Hawtrey, R. G. (1918), 'The Bank Restriction of 1797.' *Economic Journal* **28**, 52–65.

Herfindahl, O. C. (1966), 'Development of the Major Metal Mining Industries in the United States from 1839 to 1909.' In *Output, Employment, and Productivity in the United States after 1800*. New York: National Bureau of Economic Research.

Highfield, R. A., O'Hara, M., and Smith, B. (1996), 'Do Open Market Operations Matter? Theory and Evidence from the Second Bank of the United States.' *Journal of Economic Dynamics and Control* **20**, 479–519.

Holdsworth, J. T., and Dewey, D. R. (1910), *The First and Second Banks of the United States*. (Senate Document No. 571, 61st Congress, 2nd session). Washington, DC: U.S. National Monetary Commission.

Hoover, E. D., and Taylor, G. R. (1959), 'Statements and Source Materials on Prices.' In *Employment, Growth, and Price Levels*. Washington, DC: Hearings before the Joint Economic Committee, 86th Congress, 1st session.

Imlah, A. H. (1958), *Economic Elements in the Pax Britannica*. Cambridge, MA: Harvard Univ. Press.

Joines, D. H. (1985), 'Deficits and Money Growth in the United States: 1872–1983.' *Journal of Monetary Economics* **16**, 329–351.

Kemp, D. S. (1975a), 'A Monetary View of the Balance of Payments.' Federal Reserve Bank of St. Louis *Review* **57**, 14–22.

Kemp, D. S. (1975b), 'Balance-of-Payments Concepts: What Do They Really Mean?' Federal Reserve Bank of St. Louis *Review* **57**, 14–23.

Kemp, D. S. (1976), 'U.S. International Trade and Financial Developments in 1976.' Federal Reserve Bank of St. Louis *Review* **58**, 8–14.

Kindahl, J. K. (1961), 'Economic Factors in Specie Resumption: the United States, 1865–79.' *Journal of Political Economy* **69**, 30–48.

Laurent, R. D. (1974), 'Currency in Circulation and the Real Value of Notes.' *Journal of Money, Credit and Banking* **6**, 213–226.

Leacy, F. H. (Ed.) (1983), *Historical Statistics of Canada*. Ottawa: Statistics Canada (2nd edn).

Martin, D. A. (1973), '1853: The End of Bimetallism in the United States.' *Journal of Economic History* **33**, 825–844.

Martin, D. A. (1976), 'U.S. Gold Production Prior to the California Gold Rush.' *Explorations in Economic History* **4**, 437–449.

Meltzer, A. H. (1986), 'Some Evidence on the Comparative Uncertainty Experienced under Different Monetary Regimes.' In C. D. Campbell and W. R. Dougan (Eds.), *Alternative Monetary Regimes*. Baltimore, MD: Johns Hopkins Univ. Press.

Metcalf, C., Redish, A., and Shearer, R. (1996), *New Estimates of the Canadian Money Stock: 1871–1967* (Discussion Paper no. 96–17). Vancouver: Univ. of British Columbia, Department of Economics.

Miron, J. A. (1989), 'The Founding of the Fed and the Destabilization of the Post-1914 US Economy.' In M. De Cecco and A. Giovannini (Eds.), *A European Central Bank*? Cambridge, UK: Cambridge Univ. Press.

Mitchell, B. R. (1988), *British Historical Statistics*. Cambridge, UK: Cambridge Univ. Press.

Mitchell, W. C. (1908), *Gold, Prices, and Wages under the Greenback Standard*. Berkeley: Univ. of California Press.

North, D. C. (1960), 'The United States Balance of Payments, 1790–1860.' In W. N. Parker (Ed.), *Trends in the American Economy in the Nineteenth Century*. Princeton, NJ: Princeton Univ. Press.

Officer, L. H. (1996), *Between the Dollar-Sterling Gold Points*. Cambridge, UK: Cambridge Univ. Press.

Perkins, E. J. (1994), *American Public Finance and Financial Services, 1700–1815*. Columbus: Ohio State Univ. Press.

Redlich, F. (1968), *The Molding of American Banking*. New York: Johnson.

Report [1819] (1968), Appendix to Second Report. In *Reports from Secret Committees of the House of Commons and the House of Lords on the Expediency of the Resumption of Cash Payments* (British Parliamentary Papers, Monetary Policy, General, vol. 2). Shannon, Ireland: Irish Univ. Press.

Report [1832] (1968), Appendix. In *Report from the Committee of Secrecy on the Bank of England Charter* (British Parliamentary Papers, Monetary Policy, General, vol. 4). Shannon, Ireland: Irish Univ. Press.

Report [1840] (1968), Appendix. In *Report from the Select Committee on Banks of Issue* (British Parliamentary Papers, Monetary Policy, General, vol. 5). Shannon, Ireland: Irish Univ. Press.

Report [1841] (1968), Appendix. In *First and Second Reports from the Select Committee of Secrecy on Banks of Issue* (British Parliamentary Papers, Monetary Policy, General, vol. 6). Shannon, Ireland: Irish Univ. Press.

Report [1848] (1968), *Appendix and Index to the Reports from the Secret Committee on Commercial Distress*, Part 2 (British Parliamentary Papers, Monetary Policy, Commercial Distress, vol. 2). Shannon, Ireland: Irish Univ. Press.

Report [1857] (1969), *Appendix and Index to the Report from the Select Committee on Bank Acts* (British Parliamentary Papers, Monetary Policy, General, vol. 8). Shannon, Ireland: Irish Univ. Press.

Rockoff, H. (1971), 'Money, Prices, and Banks in the Jacksonian Era.' In R. W. Fogel and S. L. Engerman (Eds.), *The Reinterpretation of American Economic History*. New York: Harper & Row.

Rockoff, H. (2000), 'Review Essay: Milton Friedman and Anna Jacobson Schwartz, *A Monetary History of the United States, 1867–1960*.' [Online]. EH.Net *Review*. URL: http://www.eh.net/ bookreviews/library/rockoff.shtml.

Romer, C. D. (1989), 'The Prewar Business Cycle Reconsidered: New Estimates of Gross National Product, 1869–1908.' *Journal of Political Economy* **97**, 1–37.

Rutner, J. L. (1974), 'Money in the Antebellum Economy: Its Composition, Relation to Income, and its Determinants.' Unpublished Ph.D. dissertation, Univ. of Chicago.

Seaman, E. C. (1852), *Essays on the Progress of Nations*. New York: Scribner.

Secretary of the Treasury (various years), *Annual Report*.

Shetler, D. (1973), 'Monetary Aggregates Prior to the Civil War: A Closer Look.' *Journal of Money, Credit and Banking* **5**, 1000–1006.

Shrigley, I. (1935), *The Price of Gold*. London: King.

Shultz, W. J., and Caine, M. R. (1937), *Financial Development of the United States*. New York: Prentice Hall.

Simon, M. (1960), 'The United States Balance of Payments, 1861–1900.' In W. N. Parker (Ed.), *Trends in the American Economy in the Nineteenth Century*. Princeton, NJ: Princeton Univ. Press.

Smith, W. B. (1953), *Economic Aspects of the Second Bank of the United States*. Cambridge, MA: Harvard Univ. Press.

Solomou, S., and Weale, M. (1991), 'Balanced Estimates of UK GDP 1870–1913.' *Explorations in Economic History* **28**, 54–63.

Solomou, S., and Weale, M. (1996), 'UK National Income, 1920–1938: The Implications of Balanced Estimates.' *Economic History Review* **49**, 101–115.

Stevens, E. J. (1971), 'Composition of the Money Stock Prior to the Civil War.' *Journal of Money, Credit and Banking* **3**, 84–101.

Studenski, P., and Krooss, H. E. (1963), *Financial History of the United States*. New York: McGraw–Hill.

Sylla, R. (1982), 'Monetary Innovation and Crises in American Economic History.' In P. Wachtel (Ed.), *Crises in the Economic and Financial Structure*. Lexington, MA: Heath.

Taus, E. R. (1943), *Central Bank Functions of the United States Treasury, 1789–1941*. New York: Columbia Univ. Press.

Temin, P. (1969), *The Jacksonian Economy*. New York: Norton.

The Economist (various issues).

Timberlake, R. H. (1993), *Monetary Policy in the United States*. Chicago: Univ. of Chicago Press.

Treasury Department (1915), *Information Respecting United States Bonds, Paper Currency, and Coin*. Washington, DC: Government Printing Office.

Urquhart, M. C. (1993), *Gross National Product, Canada, 1870–1926*. Montreal: McGill–Queen's Univ. Press.

Wallich, H. C. (1950), *Monetary Problems of an Export Economy*. Cambridge, MA: Harvard Univ. Press.

Warburton, C. (1962), 'Monetary Disturbances and Business Fluctuations in Two Centuries of American History.' In L. B. Yeager (Ed.), *In Search of a Monetary Constitution*. Cambridge, MA: Harvard Univ. Press.

Warren, G. F., and Pearson, F. A. (1935), *Gold and Prices*. New York: Wiley.

Weiss, T. (1992), 'U.S. Labor Force Estimates and Economic Growth, 1800–1860.' In R. E. Gallman and J. J. Wallis (Eds.), *American Economic Growth and Standards of Living before the Civil War*. Chicago: Univ. of Chicago Press.

Wettereau, J. O. (1937), 'New Light on the First Bank of the United States.' *Pennsylvania Magazine of History and Biography* **61**, 263–285.

Wettereau, J. O. (1985), *Statistical Records of the First Bank of the United States*. New York: Garland.

Weymark, D. M. (1995), 'Estimating Exchange Market Pressure and the Degree of Exchange Market Intervention for Canada.' *Journal of International Economics* **39**, 273–295.

Weymark, D. M. (1997a), 'Measuring Exchange Market Pressure and Intervention in Two Interdependent Economies: A Two-Country Model.' *Review of International Economics* **5**, 72–82.

Weymark, D. M. (1997b), 'Measuring the Degree of Exchange Market Intervention in a Small Open Economy.' *Journal of International Money and Finance* **16**, 55–79.

Weymark, D. M. (1998), 'A General Approach to Measuring Exchange Market Pressure.' *Oxford Economic Papers* **50**, 106–121.

Woodward, G. T. (1995), 'Interest-Bearing Currency: Evidence from the Civil War Experience.' *Journal of Money, Credit and Banking* **27**, 927–937.

12 The quantity theory in New England, 1703–1749

New data to analyze an old question*

1. Introduction

The monetary system of colonial times has long fascinated scholars, primarily because of the colonies' experience with paper money. Each colony issued its own notes, on initiation of its legislature. Perhaps the most controversial issue that historians have explored and argued about in this connection is why certain colonies had stable fiat currencies and others did not. Arguments have raged about the extent to which (1) colonial exchange rates were fixed versus floating, (2) 'backing' mechanisms for paper money superseded the quantity theory, and (3) specie outflows offset paper-money issues – with the three issues related.

Rather than explore these issues directly, this paper is concerned with their underpinning. The path-breaking study of West (1978) was the first systematic quantitative investigation of the quantity theory in colonial America, and the findings of this work have been generally confirmed. West showed that there is a strong positive relationship between the price level and public paper money (both expressed in logarithms) for New England (NE) in the first half of the 18th century, and so concluded that the classical quantity theory of money (CQTM) holds for that time and place, which finding is widely accepted in the literature. That is the exception rather than the rule for colonial America. Everywhere else, prices and paper money have been found (both by West and later scholars) generally to be unrelated.

The NE exception to the rule might be a specious outcome, founded on spurious regression (emanating from nonstationarity, a matter unexplored by West) or spurious data (decidedly a concern of West, who doubts the quality of his own data). West's price index pertains only to wheat and molasses, and the money measure counts only public paper currency, excluding specie and private notes. It is possible that a relative price shift between 'wheat and molasses' and the rest of consumption or GDP just happens to coincide with a change in the paper-money measure that might actually be offset by movements in excluded monies. It is also possible that nonstationarity of the variables exacerbates the spurious outcome.

In this paper, the robustness of the West finding for NE is tested by (1) constructing alternative measures of price and the money stock, and (2) explicitly guarding against spurious regression. Does the CQTM hold for all or just some data alternatives? Further, Milton Friedman's quantity theory of money (FQTM) is given an interpretation complementary to the CQTM and is also tested for NE using alternative data while accounting for non-stationarity. A data outcome from the testing of the theories is an annual set of the equation of exchange for NE over 1703–1749, tabulated for future research.[1]

2. Economic environment

In the first half of the 18th century, the four NE colonies – Massachusetts, Rhode Island, Connecticut, and New Hampshire – faced several obstacles to economic development.[2] NE lacked both a staple crop (such as sugar or tobacco) and a large slave population, that loomed large in the economies of some other British American colonies. The land/labor ratio and average farm size fell, especially after about quarter-century, as population grew rapidly. Finally, NE participated in many of England's North American wars, that put additional pressure on resources.

Yet NE had compensating advantages. The Puritan ethic of many inhabitants fostered hard work and education. Women were an available source of labor for the family farm, that was the fundamental producing unit in early industrialization. A stable government that fostered transportation networks and a legal system that enforced contracts were put in place. Lacking an obvious export crop, merchants became entrepreneurs. They developed agriculture, fishing, whaling, forestry, shipbuilding, and simple manufacturing as export industries. Boston had a natural harbor, and the export of transportation and distribution services became an important part of the NE economy.

Fostering specialization along comparative-advantage lines, NE households demanded the goods available in England, whence most of the settlers originated. So the consumption pattern of NE, abetted by imported manufactures, was close to that of the home country. It is with reason that Newell (1998, p. 96) refers to 'New England's import–export economy.'[3]

A problem that NE governments faced was a lack of money, both to finance war-related deficit spending and to stimulate commerce. There were no mines for precious metals in NE (or the other colonies), nor was there provision by the home government for circulation of British coin in the colonies. While NE obtained specie – predominantly the Spanish (silver) dollar – from balance-of-payments surpluses with the West Indies and other areas, the specie was drained by NE's chronic deficit with England.[4] By 1718 no specie was left in circulation.

The answer was paper currency ('bills of credit') issued by the colonial governments via legislation, beginning with Massachusetts in 1690.[5] The notes either directly financed government expenditure or were lent to citizens on the security of mortgaged land. Correspondingly, future taxes or loan repayment were to be used to retire the bills. Though not redeemable on demand, the bills were in principle either tax-anticipation notes or receipts representing claims on mortgaged property.[6] However, over time the bills became pure fiat money, as NE legislatures stopped imposing sufficient taxes to retire the outstanding paper on a steady schedule, and private borrowers likewise ceased paying the interest and repaying the principal on their outstanding loans. The 'backing' mechanisms (via taxes, loan repayments, or land auction on defaulted loans) were not observed, with the ultimate result of runaway inflation – suggestive of the CQTM, at least for the late 1740s.

The colonies each had their own unit of account, with British nomenclature (pounds, shillings, and pence) but not British value; and bills of credit were denominated in the colony's unit of account. On occasion between 1737 and 1743, the NE governments issued bills of 'new tenor,' worth a stipulated multiple of an old-tenor note. However, accounting and pricing in the region remained in old-tenor terms.

Throughout the first half of the 18th century, the four NE colonies constituted a unified monetary area.[7] Bills of credit issued by any of the four colonial governments circulated throughout the region at par, the currency value of specie was also integrated, and the disappearance of coin from circulation occurred simultaneously in the four colonies. Legal values of the units of account tended to be uniform throughout the region, with the other colonies following the lead of Massachusetts. Further, exchange rates were unified, and it must have been that commodity markets were as integrated as transportation and communication permitted.

Paper currency was issued not only by the colonial treasuries but also by private institutions called 'banks' – a misnomer, as their activities were limited to note issuance and retirement. In all but one case (Boston Merchants), their operation was cut short by legislation.[8] The first such 'bank' was the New London Society for Trade and Commerce, which in 1732–1733 issued to its members notes similar to Connecticut bills of credit, on the security of mortgaged land; but the Connecticut legislature ordered the Society mortgagers to redeem their notes.

On November 30, 1733, 10 Boston merchants issued £110,000 in promissory notes, to be redeemed in silver at 19s. per ounce (3/10th, 3/10th, and 4/10th) by December 30 (1736, 1739, and 1743). In the (1736, 1739) call-in, the notes were reissued at (7/10th, 4/7th) their face value. The notes were issued by being lent at 6% interest per year, with repayment 1/10th annually in specie (to be used for redemption). The notes traded in the market at a premium over bills of credit, and after a while a high proportion of the notes was hoarded for redemption.

In imitation, New Hampshire merchants issued notes on December 25, 1734; but the Massachusetts legislature passed an act forbidding circulation of the notes in the colony, and the notes became worthless.

In 1740, the Land Bank and Silver Bank were chartered by the Massachusetts legislature, and both issued notes. The Parliamentary Bubble Act of 1720 had stated that joint-stock companies could not be formed or engage in business transactions without specific authority. In 1741, the Act was extended to the American colonies, leading both banks to redeem their notes.[9]

On January 28, 1749, the Massachusetts legislature passed an act – approved by the home government on June 28 – to redeem its bills of credit and move to a silver standard after March 31, 1750. After that date, Massachusetts residents were forbidden to deal in bills of credit of the other colonies. Specie to redeem the bills, obtained from the British Parliament in reimbursement of the colony's war expenses, arrived in Boston on September 18, 1749. Thus institution of a specie standard and breakup of the NE monetary area in the following year was assured.

3. Classical quantity theory

The CQTM tested by West (1978) embodies two propositions: (1) There is a strong positive association between the price level and the money stock. (2) The money supply is an exogenous variable. West imposes (2), which determines the direction of minimization of the sum of squared errors in econometric estimation. He transforms the variables into logarithms, which yield the elasticity of price with respect to the money stock as the crucial coefficient.[10]

West's price variable is the average of the Boston price of a bushel of wheat and a gallon of molasses (data from Cole, 1938, p. 117). This measure has the advantage of embodying actual prices within NE but the disadvantage of limitation to two goods and that are agricultural: the price measure could easily diverge from general price movements. Also, the molasses price is available only from 1720; for 1703–1749 regressions, West uses the price of wheat alone. In the present paper, the price of wheat is linked to the wheat-molasses price on the basis of the 1720 overlap, resulting in a superior price variable for the 1703–1749 period. This West price variable is converted to an 'index number' with 1703 = 1, a procedure applied to all variables in the present study.

A new price-level measure is offered as an alternative: the product of (1) the exchange rate (Massachusetts-£s per £-sterling), often used by itself to proxy price, and (2) an index of the British price level; and is shown in Table 12.1.[11] This purchasing-power-parity (PPP) variable is better at capturing 'general' price movements less tainted by relative-price shifts; but has the disadvantage that deviations from PPP can be lengthy and stubbornly persistent, as modern data show (Rogoff, 1996). The PPP price may not be as good as the West variable as a short-run measure. Additionally, international transportation costs jumped by about 30 (100) percent from 1713–1740 to the early (late) 1740s, which could have slowed down convergence to PPP – again

Table 12.1 Components of equation of exchange, and population: New England, 1703–1749 (index numbers, 1703 = 1)

Year	*Price level*	*Money supply*[a]	*Real income*	*Velocity*	*Population*
1703	1.0000	1.0000	1.0000	1.0000	1.0000
1704	0.9778	1.0290	1.0512	0.9989	1.0218
1705	0.9442	1.0527	0.7922	0.7106	1.0441
1706	1.0102	1.0258	0.6920	0.6815	1.0669
1707	1.0450	1.0274	1.3169	1.3395	1.0901
1708	1.1647	1.0413	1.3028	1.4571	1.1139
1709	1.3435	1.0119	1.1400	1.5137	1.1382
1710	1.2879	1.0696	1.0152	1.2224	1.1630
1711	1.1490	1.1976	1.2034	1.1546	1.2099
1712	1.1027	1.5589	1.1218	0.7935	1.2587
1713	1.1498	1.5354	1.2579	0.9419	1.3094
1714	1.1073	1.3460	1.2842	1.0565	1.3622
1715	1.1906	1.3758	1.7402	1.5060	1.4171
1716	1.2156	1.4436	1.4616	1.2308	1.4743
1717	1.2710	1.7852	1.4932	1.0632	1.5337
1718	1.4655	1.6765	1.5459	1.3514	1.5956
1719	1.6173	1.6190	1.4649	1.4633	1.6599
1720	1.6405	1.5471	1.4737	1.5627	1.7268
1721	1.6185	1.7118	1.3608	1.2867	1.7688
1722	1.6602	2.0081	1.4858	1.2283	1.8119
1723	1.7859	2.1523	1.9086	1.5837	1.8560
1724	2.0019	2.3101	1.9094	1.6547	1.9012
1725	2.2601	2.4914	2.1837	1.9810	1.9474
1726	2.2275	2.6560	2.1111	1.7705	1.9948
1727	2.3182	2.5501	2.1035	1.9122	2.0434
1728	2.4628	2.8985	2.1000	1.7844	2.0931
1729	2.4238	2.8252	1.7535	1.5043	2.1441
1730	2.4786	2.7579	2.1908	1.9690	2.1963
1731	2.3342	2.6835	1.9796	1.7220	2.2603
1732	2.3012	2.9240	2.4449	1.9242	2.3262
1733	2.4328	3.0354	2.2011	1.7641	2.3940
1734	2.5442	4.6869	2.0962	1.1379	2.4638
1735	2.5998	4.7119	2.4698	1.3627	2.5356
1736	3.0817	4.2583	2.8099	2.0335	2.6095
1737	3.6805	4.1784	2.8953	2.5502	2.6856
1738	3.5186	4.2814	2.7534	2.2628	2.7639
1739	3.7501	4.7561	2.9292	2.3096	2.8444
1740	4.3684	4.6061	2.8000	2.6555	2.9274
1741	4.1779	6.1590	3.0932	2.0982	2.9917
1742	4.0019	6.1732	2.5184	1.6326	3.0574
1743	3.7499	6.1661	3.0429	1.8505	3.1245
1744	4.0913	6.4048	2.5841	1.6507	3.1932
1745	4.6133	10.3876	2.4417	1.0844	3.2633
1746	4.7357	16.4967	3.4257	0.9834	3.3350
1747	6.6691	21.5369	3.5293	1.0929	3.4082
1748	7.0016	22.5515	3.2082	0.9961	3.4831
1749	7.8071	22.6140	3.9551	1.3654	3.5596

Table 12.1 Continued

Price level. Derived as a residual from a purchasing-power-parity calculation. Original data from McCusker (1978) and O'Brien (1985), for exchange rate and British price index, respectively. See Section 3 of text and Appendix A for details of construction.

Money supply. Sum of public notes, private notes, and silver in circulation. Specifically, column (4) of Table 12.5, converted to 1703 = 1. Original data on public notes from Brock (1975, 1992) and MacInnes (1952), on private notes and silver from various sources. See Section 3 of text and Appendix A for details of sources and construction.

Real income. Aggregate, not per-capita, income. Derived from assumptions of (1) trend real income proportional to population, and (2) current/trend ratio of real income equal to current/trend ratio of trade-volume with England. Original data from Bureau of the Census (1975). See Section 4 and Appendix A for details of construction.

Velocity. Constructed residually to make the equation of exchange hold: Velocity = (Price level × Real income)/(Money supply).

Population. Constructed from decadal-year figures via logarithmic interpolation. Decadal data from Bureau of the Census (1975). See Appendix A.

a End of May.

a problem not applicable to the West variable.[12] The West and PPP price series are about the same in 1745 (4.75 versus 4.61); but thereafter West increases to 11.31 in 1749, and PPP only to 7.81. Which series is more accurate?

Fortunately, there exists an empirical test of the reliability of the two price measures that incorporates the late 1740s. Brock (1992, p. 10) provides the price of two different market baskets of consumables in Massachusetts – the first for 1747 relative to 1707, the second for 1747 relative to 1717. The first index number is 7.41, compared to 6.38 for the PPP variable and 5.63 for the West variable; the second is 5.15, compared to 5.25 for the PPP variable and 5.67 for West. This comparison suggests that the PPP variable, notwithstanding its specific biases, may be more reliable than the West variable; but both are retained to see if the particular price variable 'matters' in CQTM testing.

West chooses as his money-stock variable the Brock (1975, pp. 591–592) series of NE bills of credit outstanding. Minor corrections are made to the series (PU_B) here.[13] The CQTM will be tested using various alternative measures of the money stock, in addition to PU_B. Brock's series counts Massachusetts bills at May 31, but bills of the other colonies at end-of-year. A series (PU_{BC}) with consistent timing of May 31 for all colonies is developed here and listed in column (1) of Table 12.5, in Appendix A.

For a comprehensive paper-currency series, private notes in circulation (PR) are added to PU_{BC}. The purpose is to see if their inclusion 'matters' in the testing of the CQTM. Whatever could be found or reasonably conjectured on outstanding private notes is incorporated in an end-of-May series (listed in column (2) of Table 12.5), on the assumption that these notes (except for Boston Merchants' notes) circulated at par.[14] Boston Merchants' notes are valued according to an estimated market premium over public bills. Following Brock (1992, p. 10), all outstanding Boston Merchants' notes are deemed to be in circulation until the end of 1735 and an arbitrary one-third proportion thereafter.[15]

For a conceptually complete money-stock series, specie in circulation (SI) should be included in the money stock. Again the purpose is to ascertain if silver added to public bills (PU_{BC} + SI) or to total notes (PU_{BC} + PR + SI) makes a difference in testing the CQTM. Unfortunately, there exist no quantitative data on the specie stock in NE (or any colony). All that one has to work with are guesses or conjectures by contemporary and later writers, including the important consensus that there was a steady loss of silver in the first quarter of the 18th century until the specie was all gone. Consistent specific-year conjectures of silver in circulation are accepted, and interpolation is used for the intervening years. Interpolation based on the NE balance-of-trade deficit rather than mere linearity is applied for 1703–1710.[16] The estimated silver in circulation (SI) is shown in column (3) of Table 12.5; and total money supply (PU_{BC} + PR + SI) in old-tenor pounds in column (4) of Table 12.5 and in index-number form in Table 12.1.[17]

While the estimate of silver in circulation is crude, it does mark an advance over the implicit series of Davis (1900, frontispiece) and Brock (1975, pp. 30e, 30i). Given the high proportion of silver relative to notes in the early part of the century and the steady decline in this proportion (see Table 12.5) – which, at least qualitatively, is the consensus of scholars of colonial NE – omission of silver from the money stock cannot be justified on the grounds that silver in circulation was proportional to notes or that silver was an unimportant component of the money supply. An alternative procedure is to begin the sample period in the earliest year in which it is known that all silver was gone from circulation; but the sample size would thereby be reduced substantially (by 32%, according to Table 12.5), unwise when one begins with a sample size that is not large (47 observations) and one plans to apply time-series analysis.

Prior to selection of the appropriate model to use for time-series analysis, it is necessary to determine whether the pertinent variables are stationary or nonstationary. It is true that one should not expect price indexes and money-stock indexes to be stationary or even trend-stationary; but transformation to logarithms could result in stationarity. All the price ($\log P$) and money-stock ($\log M$) series experience a sharp jump from 1744 to 1745, a structural break associated with wartime. The Perron test is applied to test a jump in a unit-root process versus a change in the intercept of a trend-stationary process. It turns out that, at the 5% level, a unit root cannot be rejected for both the West and PPP $\log P$ variables, whereas a unit root is rejected for all five $\log M$ series.

West's equation to test the CQTM is $\log P_t = \alpha + \beta \cdot \log M_t + u_t$. With $\log P$ and $\log M$ integrated of different orders, both ordinary least-squares and the cointegration model are inapplicable. Moreover, the traditional technique of first-differencing either violates the theory (if $\log P$ alone is differenced) or discards information (if both $\log P$ and $\log M$ are differenced). West applies the AR(1) model: $u_t = \rho \cdot u_{t-1} + e_t$, which corrects for nonstationarity, providing $-1 \leq \rho \leq 1$ and e_t is white noise. He uses an old technique (Cochrane–Orcutt) for estimation. Non-linear least squares is adopted here, because that method estimates α, β simultaneously with ρ rather than conditional on ρ.

Table 12.2 West model: results

Money stock		*Autoregressive coefficient*	*Corrected R^2*
Variable	*Elasticity*		
West price variable			
PU_B	0.51 (6.04)	0.72 (9.44)	.92
PU_{BC}	0.53 (6.27)	0.73 (9.93)	.92
PU_{BC} + SI	0.68 (15.57)	0.44 (3.14)	.94
PU_{BC} + PR	0.53 (6.44)	0.73 (10.19)	.92
PU_{BC} + SI + PR	0.67 (16.94)	0.41 (2.88)	.94
PPP price variable			
PU_B	0.48 (8.85)	0.73 (13.26)	.96
PU_{BC}	0.47 (8.75)	0.73 (12.88)	.96
PU_{BC} + SI	nonstationary process		
PU_{BC} + PR	0.46 (8.55)	0.73 (12.10)	.96
PU_{BC} + SI + PR	nonstationary process		

Dependent variable: Price level.
PU_B, public notes (Brock series).
PU_{BC}, public notes (time-consistent Brock series).
PR, private notes.
SI, silver.
All variables in logarithms.
t statistics in parentheses.
Time period: 1703–1749.

Table 12.2 provides results of estimating the West model for the alternative price and money-stock variables. The estimate of the elasticity β and the associated t value are in the second column, followed by the estimate of ρ with its t value. The equations show that inclusion of private notes (PR) makes little difference to the estimates.[18] In contrast, including silver (SI) in the money stock increases both the size and significance of the estimated elasticity. The problem is that, for the PPP-price equations involving SI, the AR(1) process is nonstationary, and, further, the error e_t for all West-price equations except PU_B is not white noise: there is a significant autocorrelation at lag two.[19] In sum, one cannot reach conclusions about the validity of the CQTM, because most regressions could be spurious.

An alternative procedure is to include $\log P_L$ (where 'L' denotes a one-period lag) as an additional explanatory variable. This formulation has the advantage of focusing on the nonstationary variable itself rather than on the error term. Also, past effects of the money stock on price are embedded in lagged price. This means that the elasticity coefficient is conditional on lagged price, and will surely be lower than the corresponding unconditional coefficient of the West model. Results using ordinary least-squares are presented in Table 12.3. Residuals for all the West-price equations have a spike at lag

Table 12.3 Conditional-elasticities model: results

Money stock		*Lagged-price coefficient*	*Corrected R^2*
Variable	*Elasticity*		
West price variable			
PU_B	0.07 (1.74)	0.89 (9.84)	.93
PU_{BC}	0.07 (1.83)	0.88 (9.78)	.93
PU_{BC} + SI	0.37 (5.24)	0.50 (4.78)	.95
PU_{BC} + PR	0.08 (1.91)	0.87 (9.67)	.93
PU_{BC} + SI + PR	0.39 (5.64)	0.47 (4.48)	.95
PPP price variable			
PU_B	0.04 (1.66)	0.95 (18.59)	.98
PU_{BC}	0.04 (1.74)	0.95 (18.60)	.98
PU_{BC} + SI	0.13 (3.24)	0.84 (13.44)	.98
PU_{BC} + PR	0.04 (1.65)	0.95 (18.30)	.98
PU_{BC} + SI + PR	0.13 (3.10)	0.84 (12.76)	.98

Dependent variable: Price level.
PU_B, public notes (Brock series).
PU_{BC}, public notes (time-consistent Brock series).
PR, private notes.
SI, silver.
All variables in logarithms.
t statistics in parentheses.
Time period: 1703–1749.

two; so those regressions could be spurious. However, residuals for the PPP-price equations are clean. Again, private notes (PR) have little impact; but inclusion of silver (SI) in the money supply enhances elasticity and provides good support for the CQTM.

4. Friedman quantity theory

Milton Friedman's restatement of the quantity theory is generally construed as a theory of the demand for money and as an alternative to the CQTM. A different interpretation is offered here, for which one must return to Friedman's (1956) original statement of the quantity theory. Friedman restated the equation of exchange with velocity the key variable; and velocity is neither constant nor unknown, but rather a stable function of real income and other variables (expected inflation, returns on financial assets, wealth, and tastes). These other variables do not include the price level or the money stock. Therefore, Friedman can be interpreted as taking the real money supply (M/P) and dividing it in another way (Y/V). Just as the CQTM is concerned with the effect of M on P, so the FQTM is with the effect of Y (and other variables, but excluding the levels of M and P) on V. Friedman's theory is thus interpreted here as both symmetrical to, and complementary to, the classical theory.

Directly measured data on colonial income, whether nominal or real, do not exist. Like other researchers, one can use only a crude technique to measure *real* income, required here to test the FQTM. Assume that in the long run, per-capita real income, is constant, that is, trend output is proportional to population. Many historians believe that this relationship holds broadly for the colonial period.[20] There exists specific empirical evidence that it is a good approximation for NE in 1703–1749.[21] For the short run, it is assumed that the annual current/trend output ratio is equal to the current/trend trade-volume (with England) ratio. The result is a real-income (Y) series in index-number form for 1703–1749, listed in Table 12.1. Velocity (V) is computed in the usual way, to make the equation of exchange hold: $V = (P \cdot Y)/M$, and presented in Table 12.1 using the PPP price variable. An alternative velocity series is constructed using the West price variable.

Because the implied estimate of *nominal* income ($P \cdot Y$) is an index number, so too is velocity (V). Fortunately, to test the quantity theories, it suffices to have components of the equation of exchange in index-number form.

It must be emphasized that real income (Y), shown in Table 12.1, is an aggregate rather than per-capita measure. Further, deliberately, the series exhibits zero long-run per-capita growth, which is clearly shown by comparison with an index of NE population (final column of Table 12.1). Although the real-income series for this study is constructed on the assumption of zero long-run per-capita growth (that is, trend aggregate income proportionate to population), on the grounds stated in notes 20–21; absence of per-capita growth is not an essential element of the technique. A nonzero (positive or negative) growth of trend per-capita real income is readily incorporated, as shown in Appendix A.

While the technique to estimate income accounts for cycles as well as trend, short-run movements in trade volume may not well match short-run movements in total real income. For example, war may alter short-run amplitude of trade over trend more than war's impact on total real income. This bias, as well as any biases involving the price and money-stock variables, carries over into the velocity variable; although if one is lucky, there could be some cancellation of measurement errors.

For NE in 1703–1749, usable data on wealth do not exist, and interest rates were essentially constant.[22] Tastes (which incorporate technology) may be taken as unvarying for this period. Expected inflation is typically proxied by actual inflation $\{\Delta \log P = \log(1 + \Delta P/P_L)$, where $\Delta P/P_L$ is the conventional inflation rate$\}$; but this variable should incorporate the fact that, from 1747 onward, it was expected that there would be a currency reform and return to a silver standard in Massachusetts.[23] So the effect of inflation is altered by inclusion also of the variable $(D \cdot \Delta \log P)$, where $D = 0$ in 1703–1746, 1 in 1747–1749. Nevertheless, the measure of inflation is crude and likely biased.

The augmented Dickey–Fuller test for 1703–1749 shows that $\log Y$ is trend-stationary (perhaps, in part, an outcome of how Y is constructed), while a unit root cannot be rejected for both log$V$ series (none of which variables experiences the 1744–1745 jump of the log$P$ and log$M$ series). Thus, symmetrical with the CQTM, the dependent variable for the FQTM is found to be nonstationary and the principal explanatory variable stationary. In response, the CQTM testing indicates that logV_L be included as an additional explanatory variable. This setup has the advantage of adhering to rational expectations of inflation; for the inflation (and income) elasticities are now conditional on lagged velocity. With efficient markets, all information concerning future inflation is embedded in last period's velocity. Then it is reasonable to proxy new information regarding expected inflation via the current-period's actual inflation. Of course, the error term must be white noise – else the results could be spurious.[24]

Results are shown in Table 12.4. For both the West-price and PPP-price variable, the residuals of the equation satisfy the test for white noise. PPP-based velocity yields stronger results; but the coefficient and significance of income are about the same. Estimated elasticities are consistent with Friedman's theory. Velocity is positively related to real income, but the effect is inelastic (so real demand for money also increases with income).

Inflation, proxying expected inflation, normally increases velocity. However, in 1747–1749 inflationary expectations were altered by the plan to redeem all outstanding Massachusetts bills of credit and return to a specie standard. Now inflation leads to a net decline in velocity. This result is explainable as follows. The higher inflation, the greater the probability that the plan would be consummated, whence hoarding of bills increases (velocity falls). This could explain why, mirroring the price variable, velocity remained relatively low in the late 1740s for the PPP-based series (though it increases substantially in 1748–1749 for the West-price-based series). Nevertheless,

Table 12.4 Friedman model: results

Variable	*Coefficient*	
	West price variable	*PPP price variable*
Constant	0.07 (1.14)	—
logV_L	0.43 (2.86)	0.55 (5.21)
log$Y$	0.26 (2.92)	0.23 (3.00)
Δlog$P$	0.35 (1.70)	1.02 (2.57)
$D\cdot\Delta$log$P$	–0.54 (1.30)	–1.62 (2.41)
Corrected R^2	.49	.68

Dependent variable: log V, with V derived alternatively from West and PPP price variable.
V, Velocity.
Y, Real income.
P, Price level.
D = 0 in 1703–1746, 1 in 1747–1749.
t statistics in parentheses.
Time period: 1703–1749.

notwithstanding the crudity of the data, the FQTM model (admittedly in simple form) does receive support from both velocity measures.

5. Concluding comments

This study develops annual time series of all components of the equation of exchange for NE for 1703–1749 – the first time this has been done for the colonial experience. It is true that the data have limitations; but the price variable by construction is superior to the usual exchange-rate proxy and by empirical testing appears more reliable than the Cole price series used by West. Also, the money-stock series is distinguished by improved data for public bills of credit and Boston Merchants' notes, and, for the first time for NE, incorporation of other private notes and specie into the money supply. The income measure (and therefore velocity) is based on a short-run behavioral assumption that is an uncertain approximation to reality, but is not needed for testing of the CQTM. Applying modern time-series analysis to the data, the CQTM and FQTM are tested, and both receive support – which confirms the thrust of West's findings.

Appendix A. Data sources and construction of variables

Exchange rate The exchange rate, number of Massachusetts £s per £ sterling, is from McCusker (1978, pp. 140–141), with missing data for 1704 and 1707–1708 obtained via linear interpolation.

British price index O'Brien (1985, pp. 788–789, 793–794) provides indexes of agricultural prices and industrial prices, which need to be weighted for an aggregate index. Weights of .58 and .42 are suggested by Officer (2003, pp. 67, 83–84, n. 65), and they are employed here.

Price level The NE price level is the product of the exchange rate and British price index (both expressed as indexes 1703 = 1).

Public bills of credit Brock (1975, pp. 591–592; 1992, Table 1) is the source for outstanding Massachusetts bills and MacInnes (1952, pp. 588–590) for Rhode Island bills, end-of-May. The source for Connecticut and New Hampshire bills end-of-year is Brock (1992, Table 1). Data do not exist to obtain generally Connecticut and New Hampshire figures for other than end-of-year; so linear interpolation is used to convert the two series to end-of-May.

Boston Merchants' notes The amount of outstanding notes in face-value pounds at end-of-May 1734–1743 is as shown in column (3) of Table 12.6 A 'theoretical premium' of Boston Merchants' notes over public bills is

Table 12.5 Components of money supply, New England, 1703–1749 (old-tenor pounds)

Year[a]	*Notes in circulation*		*Silver in circulation (3)*	*Money supply*[c] *(4)*
	Public (1)	*Private*[b] *(2)*		
1703	6,431	0	173,848	180,279
1704	17,675	0	167,834	185,509
1705	29,455	0	160,329	189,784
1706	31,124	0	153,798	184,922
1707	40,825	0	144,399	185,224
1708	57,003	0	130,725	187,728
1709	67,614	0	114,802	182,416
1710	96,414	0	96,414	192,828
1711	132,560	0	83,337	215,897
1712	210,780	0	70,260	281,040
1713	219,132	0	57,676	276,808
1714	197,562	0	45,092	242,654
1715	215,513	0	32,508	248,021
1716	240,331	0	19,925	260,256
1717	314,486	0	7,340	321,826
1718	302,231	0	0	302,231
1719	291,867	0	0	291,867
1720	278,908	0	0	278,908
1721	308,592	0	0	308,592
1722	362,019	0	0	362,019
1723	388,006	0	0	388,006
1724	416,458	0	0	416,458
1725	449,154	0	0	449,154
1726	478,824	0	0	478,824
1727	459,732	0	0	459,732
1728	522,533	0	0	522,533
1729	509,326	0	0	509,326
1730	497,184	0	0	497,184
1731	483,771	0	0	483,771
1732	527,129	0	0	527,129
1733	532,322	14,904	0	547,226
1734	722,044	122,906	0	844,950
1735	705,307	144,145	0	849,452
1736	721,340	46,341	0	767,681
1737	722,208	31,073	0	753,281
1738	738,236	33,613	0	771,849
1739	823,203	34,221	0	857,424
1740	810,829	19,555	0	830,384
1741	923,502	186,837	0	1,110,339
1742	1,020,093	92,811	0	1,112,904
1743	1,092,064	19,555	0	1,111,619
1744	1,154,645	0	0	1,154,645
1745	1,872,654	0	0	1,872,654
1746	2,974,006	0	0	2,974,006
1747	3,882,648	0	0	3,882,648
1748	4,065,562	0	0	4,065,562
1749	4,076,818	0	0	4,076,818

Table 12.5 Continued

a End of May.
b 'Total' row of Table 12.7 plus, for (1734–1735, 1736–1743), (all, one-third) of column (4) of Table 12.6.
c Sum of columns (1)–(3).

Table 12.6 Derivation of Boston Merchants' notes outstanding

Year[a]	*Premium of notes over Massachusetts old-tenor bills (%)*		*Notes outstanding (pounds)*	
	Theoretical (1)	*Market (2)*	*Face-value (3)*	*Old-tenor*[b] *(4)*
1734	–4.83	11.73	110,000	122,906
1735	7.95	23.77	110,000	136,145
1736	11.30	26.38	110,000	139,023
1737	6.73	21.07	77,000	93,221
1738	17.36	30.96	77,000	100,838
1739	30.77	33.33	77,000	102,664
1740	22.25	33.33	44,000	58,665
1741	27.91	33.33	44,000	58,665
1742	35.58	33.33	44,000	58,665
1743	57.71	33.33	44,000	58,665

a End of May.
b Computed, in principle, as (1 + column (2)/100) · column (3); but column (2) shown here is rounded, whereas column (4) is not.

computed as follows, using a discount rate of 6% (the legal and market rate).[25] First, compute the present value (PV, number of ounces of silver per shilling Boston Merchants' notes), using the annual-compounding formula:

$$PV = (A/1.06^{a} + B/1.06^{b} + C/1.06^{c})/19, \tag{A.1}$$

where, for 1733–1736, $A = B = .3$, $C = .4$; for 1737–1739, $A = 0$, $B = 3/7$, $C = 4/7$; and for 1740–1743, $A = B = 0$, $C = 1$. The parameters (a, b, c) are the number of years, including fractions, to the end of (1736, 1739, and 1743), and 1/19 is the fixed redemption rate (number of ounces of silver per shilling Boston Merchants' notes). The theoretical percent premium ($\mathrm{PR_{th}}$) of Boston Merchants' notes over old-tenor Massachusetts bills is

$$\mathrm{PR_{th}} = 100 \cdot (\mathrm{PV}{\cdot}P_S - 1), \tag{A.2}$$

where P_s is the annual (representing the current) price of silver (number of old-tenor shillings per ounce; from Brock, 1992, Table 2). Note that $PV{\cdot}P_S$ has proper dimension, number of old-tenor shillings per shilling Boston Merchants' notes. Formulas (A.1) and (A.2) are used to compute $\mathrm{PR_{th}}$ at end-of-May 1734–1743, shown in column (1) of Table 12.6. It is observed that

PR_{th} increases monotonically within each redemption period (1734–1736, 1737–1739, and 1740–1743).

The theoretical premium is used to estimate the actual (market) premium, PR_{mk}. Initially after issuance, at least to the end of 1733, the notes were at par with Massachusetts bills (Anonymous, 1734; in Davis, 1911, vol. 3, p. 32; Brock, 1992, p. 9). With P_S 23 at the end of 1733, PR_{th} is computed as –16.87%, while PR_{mk} is zero for that date. At the end of 1739, PR_{mk} was 33.33%, with P_S 29 and PR_{th} 20.90%. Let excess premium, $PR_{ex} = PR_{mk} - PR_{th}$. Then PR_{ex} is (16.87, 12.43) at end of (1733, 1739). Linear interpolation is used to obtain PR_{ex} for end-of-May 1734–1739. PR_{mk} for end of 1739 also holds for the entirety of 1740 and 1741.[26] Absent other information, it is reasonable to extend this premium forward to end-of-May 1743 and backward to end-of-May 1739 (where it supersedes the interpolated value). Adding the interpolated PR_{ex} to PR_{th}, end-of-May PR_{mk} is generated for 1734–1738. Thus a PR_{mk} series for end-of-May 1734–1743 is obtained, and is shown in column (2) of Table 12.6. Outstanding Boston Merchants' notes are converted from face-value denomination to old-tenor terms by multiplying the face-value amount by $(1 + PR_{mk}/100)$. The resulting series is shown in column (4) of Table 12.6, and the appropriate circulation proportion (unity 1734–1735, one-third 1736–1743) is applied to this series.

New London Society notes Circulation at end of May 1733 is the total of notes emitted (£14,904), that occurred between October 1732 and January 1733, because retirements did not begin until later (probably October). In May 1734 total retirements reached £14,000, indicating that by May 31 circulation had ended. Information is from Stark (1984, pp. 6; 18, n. 31; 19, n. 43).

New Hampshire Merchants' notes All emissions occurred at the end of 1734, none in 1735, as indicated in the preamble to the Massachusetts Act prohibiting circulation of the notes in that colony. The bills became worthless after it was apparent that an order disallowing the Act was not readily forthcoming, that is, after 1735 (Davis, 1901, pp. 126, 129). Therefore, the notes enter the money supply only for May 31, 1735. The amount of bills issued, reported to the Commissioners of Trade in London, was £6000 sterling. It is a safe assumption that the legal maximum rate of £$1\frac{1}{3}$ New Hampshire = £1 sterling was used for conversion, whence £8000 in NE currency.[27]

Land Bank notes Issuance began in September 1740, and by May 31, 1741 had ceased, with £47,282 having been placed in circulation; for news of the Parliamentary Act thwarting the Land and Silver Banks reached Massachusetts on May 28. Calling-in of the notes began later, with £32,500 retired by June 30, 1742, leaving £14,782 in circulation.[28] Linear interpolation yields £17,282 in circulation at the end of May 1742. It is reasonable to assume that all notes had been successfully retired by end-of-May 1743.

Silver Bank notes Issuance began in August 1740, and £120,000 were in circulation at end-of-May 1741. By June 30, 1742, £69,362 had been retired, leaving £50,638 remaining in circulation.[29] Linear interpolation yields £55,974 for end-of-May 1742. Again it is a reasonable assumption that no notes were in circulation at end-of-May 1743.

Circulation of private notes, other than Boston Merchants' notes, at May 31 is listed in Table 12.7.

Silver A contemporary estimate of £200,000 in circulation is taken by Davis to apply to the beginning of 1700. He argues that there is empirical support in favor of the figure, and the figure and date are accepted by Brock without reservation.[30] Therefore the initial data point for interpolating silver in circulation is £200,000 at the end of 1699.

Two alternative years by which all silver disappeared from circulation have been offered, 1718 and 1726 – the former by Anonymous (c. 1749; in Davis, 1911, vol. 4, p. 402), Douglass (1740; in Davis, 1911, vol. 3, p. 334), and Davis (1900, p. 390); the latter by Anonymous (1743; in Davis, 1911, vol. 4, p. 157). There exists additional contemporary evidence that end of 1717 is the more likely date than end of 1725, and modern historians concur.[31] Therefore, it is reasonable to adopt end of 1717 as the earliest date at which no silver remained in circulation.

Brook's (1992, p. 7) judgment that 'in 1710 there was perhaps as much silver in circulation as there were bills' is accepted. A contemporary observation (Anonymous, 1743; in Davis, 1911, vol. 4, p. 154) that in 1712 silver in circulation was one-third of paper currency is also adopted. So silver in circulation at end-of-May (1710, 1712) is set equal to (total, one-third) paper currency in circulation.

Linear interpolation is used to generate silver in circulation for end-of-May 1711 and 1713–1717. For 1703–1709, a superior interpolation technique is employed. The annual NE trade deficit with England (as a proportion of the cumulative deficit January 1, 1700–May 31, 1710) is the interpolating factor for end-of-year 1700–1709 and end of May 1710.[32] Then linear interpolation is applied to generate end-of-May 1703–1709.

Table 12.7 Other private notes in circulation, New England, 1703–1749 (old-tenor pounds)

Organization	*1733*[a]	*1735*[a]	*1741*[a]	*1742*[a]
New London Society	14,904			
New Hampshire Merchants		8000		
Land Bank			47,282	17,282
Silver Bank			120,000	55,974
Total	14,904	8000	167,282	73,256

a End of May.

An important data characteristic is that from 1703 onward EX and IM are the sums of physical quantities valued predominantly at official *fixed* prices, and therefore represent volume rather than value of trade. In 1700–1702, in contrast, the official prices were revised to reflect changing market prices.[33] Therefore, the balance-of-trade deficit in *current* £s sterling is computed as $(IM - EX)$ for 1700–1702, $P_B \cdot (IM - EX)$ for 1703–1709, where P_B is the British price index rebased to 1700–1702 = 1.

Income Population (POP) for the NE colonies for decadal years (1700, 1710, . . ., 1750) is taken from Bureau of the Census (1975, p. 1168). Following McCallum (1992, p. 148, n. 13), logarithmic interpolation is used to obtain the intervening years. The volume (TR) of NE trade with England is $TR = EX + IM$. Smoothing TR via the Hodrick–Prescott procedure with parameter 100 results in trend trade-volume (TR_T). Letting (Y, Y_T) denote (actual, trend) real income, the estimator for Y is

$$\hat{Y} = (TR/TR_T) \cdot Y_T. \tag{A.3}$$

The problem is that Y_T is unknown. The study assumes that trend per-capita output (Y_T/POP) is constant, whence $Y_T = c \cdot \text{POP}$, c a positive constant. Then the estimator of Y is

$$Y^* = (TR/TR_T) \cdot c \cdot \text{POP} = (TR/TR_T) \cdot \text{POP}. \tag{A.4}$$

Without loss of generality, $c = 1$; because, in rebasing Y^* to 1703 = 1, c is cancelled.

Suppose one were to assume, rather, that trend per-capita real income grows by $100 \cdot x$ percent per year, where x is known: $(Y_T/\text{POP}) = (1 + x) \cdot (Y_T/\text{POP})_L$. Taking POP as a predetermined variable, $Y_T = (1 + x) \cdot (Y_{T,L}/\text{POP}_L) \cdot \text{POP}$. Let the initial (year-0) value of Y_T be proportional to POP: $Y_{T,0} = c \cdot \text{POP}_0$. Then $Y_T = (1 + x)^n \cdot c \cdot \text{POP}$, where n is the number of years since the initial year (0 = 1703). The estimator of Y becomes

$$Y^{**} = (TR/TR_T) \cdot (1 + x)^n \cdot \text{POP}, \tag{A.5}$$

where again c is taken as unity, with Y^{**} rebased. Of course, for zero trend per-capita growth (x = zero), Y^{**} reduces to Y^*. Note that x between .003 and .006 corresponds to the McCusker–Menard position (see note 20).

Notes

* Reprinted from *Explorations in Economic History*, vol. 42, Lawrence H. Officer, 'The Quantity Theory in New England, 1703–1749: New Data to Analyze an Old Question,' pp. 101–121. Copyright 2005, with permission from Elsevier.

1 The data appendix describes construction of the variables.

2 Maine was part of Massachusetts until becoming a state in 1820, as part of the

Missouri Compromise. Vermont, not a colony and which became the 14th state in 1791, was undeveloped land and had no status at the time.

3 On the economic development of colonial NE, see McCusker and Menard (1985); Newell (1998, 2000); Bailyn (2000); and Temin (2000).

4 A discussion of NE's balance of payments with various regions is in Nettels (1934, pp. 67–98).

5 However, prior to 1703, there are no data on outstanding bills of credit.

6 Discussions of the two types of bills of credit are in Smith (1984, 1985) and Perkins (1994, pp. 42–46).

7 The contemporary evidence for this is overwhelming – see Brock (1975, pp. 35–36, n. 23) and Bullock (1900, pp. 209–210) – and historians agree: 'The relevant monetary region included all of New England . . . There seems to be general agreement among scholars on this point' (McCallum, 1992, p. 149).

8 For histories of these banks, see Bullock (1900, pp. 223–224); Davis (1901); Hutchinson (1768/1936, pp. 288–289, 298–301); MacInnes (1952, pp. 210–215); Brock (1975, pp. 50–56); and Perkins (1991, pp. 17–20).

9 Even after 1741, true (deposit-creating) private banks with six or fewer principals would have been legal; the absence of such banks from the colonial economy is 'a profound mystery' (Perkins, 1994, p. 42).

10 West offers regressions also with the money stock lagged one and two periods as additional independent variables; but this procedure does not enhance results, due to multicollinearity, and is not repeated here.

11 Components of the equation of exchange are exhibited to four decimal places, not as a reflection of accuracy of the data but rather to ease replication and extension of results by scholars.

12 The data emanate from the annual freight-rate series of tobacco from Maryland to London (Shepherd and Walton, 1972, pp. 188–189).

13 The amount £8000 is added for 1740 (corrected in Brock, 1992, table 1), £6240 for 1747, and £5240 for 1748–1749 (Brock's figures for Rhode Island below his source by these amounts).

14 It is known, for example, that in 1741 the notes of the Silver Bank were quoted at par with public bills (Sumner, 1876, p. 32).

15 Brock writes that 'at least' a majority of the notes were hoarded; Perkins (1991, p. 18) judges that 'the vast majority' were hoarded. In favor of the position that a proportion of outstanding notes circulated after 1735 are the following facts: (1) The notes were emitted in low denominations – notes of 18 pence and a half-crown (2½ shillings) are extant (Davis, 1901, pp. 123–124) – and were therefore meant for circulation and must have been so used in an era of chronic shortage of low-denomination coin (Hanson, 1979, 1980). (2) The notes were clearly transacted; they had a market, with quotations stated in terms of public bills.

16 The technique is justified by the following: (1) Prior to 1711, it appears that silver was lost only by export, not by hoarding (Anonymous, 1749; in Davis, 1911, vol. 4, p. 380; Brock, 1992, p. 7). (2) While the trade data (official British statistics, in Bureau of the Census, 1975, pp. 1176–77) exclude regions other than England, NE's trade with (and specie inflow from) the West Indies, an important partner, remained stable until 1720 (Osgood, 1924, pp. 547–548, n. 5; Metz, 1945, p. 185). (3) Although NE exports (EX) are valued c.i.f. and imports (IM) f.o.b. London, and although other balance-of-payments items are not covered, it may be that *unrecorded* visible and invisible items (including, on the export side, NE ships sold to England and NE provision of transportation and distribution services) approximately balanced annually.

17 Some authors believe that 'book credit' was the principal component of the NE money stock. However, book credit eventually had to be cleared with currency

(paper or coin) or with barter goods. Barter goods by definition are not currency or even 'money.' So book credit is not separate from currency.

18 Table 12.5 shows that private notes are substantial in only 3 years (1734, 1735, and 1741), and constitute at most 17% of the money supply.

19 Also, the PPP-price equation for PU_{BC} has a spike at lag one. The conventional criterion of correlation having magnitude exceeding two standard errors ($2/T^{.5}$, where T is the number of observations) is applied.

20 Smith (1988, p. 23) states, 'Existing historical evidence suggests that in the colonies long-run variations in per capita real income and nominal interest rates were relatively minor.' Mancall and Weiss (1999) estimate a very slow, almost imperceptible, rate of growth of per-capita real GDP in colonial America. In contrast, McCusker and Menard (1985, p. 55) believe that there was significant growth in per-capita output, with a preferred rate of 0.6% per year and a lower bound of 0.3%. The Mancall–Weiss position is persuasive, because the authors deliberately adopt assumptions that bias their conjectures in favor of growth.

21 Anderson (1979) computes an index of productivity in agriculture in Hampshire County, Massachusetts, for the decades 1700–1709, 1710–1719, . . ., 1740–1749. The average of the indexes for the five decades is 99.4, with 1700–1709 = 100. Egnal (1998, pp. 51, 55) finds, regarding per-capita real income for the northern colonies, that there is 'a cycle with no overall gains between 1713 and 1745.' He sees the work of others as indicating an 'absence of secular growth' of per-capita income in Massachusetts between 1715 and 1745.

22 In 1693, the legal rate of interest in Massachusetts was reduced from 8 to 6% per year, where it remained throughout 1703–1749 (Anonymous, 1740; in Davis, 1911, vol. 3, p. 392; Weeden, 1890, p. 178). There is evidence that 6% was not only the legal maximum interest rate but also generally the effective interest rate. For example: 6% was the rate for loans of Connecticut bills of credit to replace New London Society notes, and for operations in connection with Boston Merchants' notes (Davis, 1901, pp. 115, 125). A contemporary observer refers to 6% as 'the lawful and common interest at Boston' (Anonymous, 1744; in Davis, 1911, vol. 4, p. 249).

23 Davis (1900, p. 209) states: '[From January 15, 1747] as time went on, and the conviction ripened that the application for reimbursement [from Parliament for expenses of the Louisburg expedition in 1745] was likely to prove successful, thoughtful men began to appreciate the fact that the reception in the province of so large an amount of coin would furnish an opportunity for the resumption of specie payments.' McCallum (1992, p. 151, n. 18), also, observes that 'monetary reform was anticipated from 1747 on.'

24 There also could be unaccounted simultaneity. It may not be serious, because (1) real income is considered exogenous in the long run, and (2) inflation is a rate of change, while velocity is a level variable. Further, there is the errors-in-variable problem of V calculated from Y, P, M to make the equation of exchange hold; this is a problem common to all studies of velocity.

25 Such a theoretical premium is calculated by both a contemporary observer and a modern historian. See Douglass (1738, also quoted by Fry, 1739; in Davis, 1911, vol. 3, pp. 230, 270) and Brock (1992, pp. 9–10, 25, n. 26).

26 See Douglass (1741, in Davis, 1911, vol. 4, p. 71); Summer (1876, pp. 31–32), and Davis (1901, pp. 135–136).

27 The legal maximum 'par of exchange' was computed implicitly as the ratio *of* a ceiling rate of 6 shillings colonial currency per Spanish dollar of weight $17\frac{1}{2}$ ounces *to* the official sterling rate of 4s. 6d. This legal maximum $33\frac{1}{3}$% overvaluation of foreign coin relative to sterling valuation was evaded in NE. See Bigham (1911, pp. 161–163) and Nettels (1934, pp. 243, 246, 248).

28 See Anonymous (1744; in Davis, 1911, vol. 4, p. 303); Davis (1901, p. 159); and Brock (1975, p. 54).

29 See Felt (1839, pp. 107, 114); Davis (1911, vol. 4, p. 303); and Brock (1975, p. 54).
30 See Anonymous (1721; in Davis, 1911, vol. 2, pp. 326–329); Davis (1900, p. 376; 1910, vol. 1, p. 34; 1911, vol. 2, p. 333); and Brock (1975, p. 23; 1992, p. 6). Davis' graph (1900, frontispiece) shows £200,000 for 1700 and zero for 1718, meaning the beginning of 1718 (see below). Therefore, the £200,000 data point correspondingly pertains to the beginning of 1700.
31 See Anonymous (1719; in Davis, 1910, vol. 1, p. 352); Anonymous (1721; in Davis, 1911, vol. 2, p. 329); Metz (1945, pp. 56, 65); Michener (1987, p. 294); and McCallum (1992, p. 151).
32 For January 1–May 31, 1710, the deficit is taken as 5/12th the total 1710 figure.
33 For histories of this recording during the first half of the 18th century, see Ashton (1960, pp. 2–7); Clark (1938, pp. 1–42); Davis (1979, pp. 77–80), and McCusker (1971, pp. 607–618).

References

Anderson, T. L., 1979. Economic growth in Colonial New England: 'Statistical Renaissance'. Journal of Economic History 39, 243–257.

Anonymous, 1719. The Present Melancholy Circumstances of the Province Considered. In: Davis, A. M., 1910, vol. 1, pp. 351–363.

Anonymous, 1721. The Second Part of South-Sea Stock. In: Davis, A.M., 1911, vol. 2, pp. 303–332.

Anonymous, 1734. A Letter from a Gentleman to his Friend. In: Davis, A. M., 1911, vol. 3, pp. 29–34.

Anonymous, 1740. An Inquiry into the Nature and Uses of Money. In: Davis, A. M., 1911, vol. 3, pp. 365–474.

Anonymous, 1743. An Enquiry into the State of the Bills of Credit. In: Davis, A. M., 1911, vol. 4, pp. 149–209.

Anonymous, 1744. An Account of the Rise, Progress and Consequences of the Two Late Schemes. In: Davis, A. M., 1911, vol. 4, pp. 237–349.

Anonymous, 1749. A Brief Account of the Rise, Progress, and Present State of the Paper Currency of New England. In: Davis, A. M., 1911, vol. 4, pp. 377–398.

Anonymous, c. 1749. The Deplorable State of New England. In: Davis, A.M., 1911, vol. 4, pp. 400–405.

Ashton, T. S., 1960. Introduction. In: Schumpeter, E. B. (Ed.), English Overseas Trade Statistics, 1697–1808. Clarendon Press, Oxford, pp. 1–14.

Bailyn, B., 2000. Slavery and population growth in Colonial New England. In: Temin, P. (Ed.), Engines of Enterprise: An Economic History of New England. Harvard University Press, Cambridge, MA, pp. 253–259.

Bigham, C. S. (Ed.), 1911. British Royal Proclamations Relating to America, 1603–1783. American Antiquarian Society, Worcester, MA.

Bullock, C. J., 1900. Essays on the Monetary History of the United States. Macmillan, New York.

Brock, L. V., 1975. The Currency of the American Colonies, 1700–1764. Arno Press, New York.

Brock, L. V., 1992. 'The Colonial Currency, Prices, and Exchange Rates.' Essays in History 34. University of Virginia.

Bureau of the Census, 1975. Historical Statistics of the United States, Colonial Times to 1970. Washington, DC. Part 2.

Clark, G. N., 1938. Guide to English Commercial Statistics, 1696–1782. Royal Historical Society, London.

Cole, A. H., 1938. Wholesale Commodity Prices in the United States, 1700–1861. Harvard University Press, Cambridge, MA.

Davis, A. M., 1900 and 1901. Currency and Banking in the Province of the Massachusetts Bay. Macmillan, New York, Parts 1 and 2.

Davis, A. M., 1910 and 1911. Colonial Currency Reprints. Prince Society, Boston, vols. 1–4.

Davis, R., 1979. The Industrial Revolution and British Overseas Trade. Leicester University Press, Leicester.

Douglass, W., 1738. An Essay Concerning Silver and Paper Currencies. In: Davis, A. M., 1911, vol. 3, pp. 217–250.

Douglass, W., 1740. A Discourse Concerning the Currencies of the British Plantations in America. In: Davis, A. M., 1911, vol. 3, pp. 307–356.

Douglass, W., 1741. A Letter to_Merchant in London. In: Davis, A. M., 1911, vol. 4, pp. 67–80.

Egnal, M., 1998. New World Economies. Oxford University Press, New York.

Felt, J. B., 1839. Historical Account of Massachusetts Currency. Perkins & Marvin, Boston.

Friedman, M., 1956. The quantity theory of money – a restatement. In: Friedman, M. (Ed.), Studies in the Quantity Theory of Money. University of Chicago Press, Chicago, pp. 3–21.

Fry, R., 1739. A Scheme for a Paper Currency. In: Davis, A. M., 1911, vol. 3, pp. 255–277.

Hanson II, J. R., 1979. Money in the Colonial American economy: an extension. Economic Inquiry 17, 281–286.

Hanson II, J. R., 1980. Small notes in the American Colonies. Explorations in Economic History 17, 411–420.

Hutchinson, T., 1768/1936. The History of the Colony and Province of Massachusetts-Bay, vol. 2. Harvard University Press, Cambridge, MA.

MacInnes, J. B., 1952. Rhode island bills of public credit, 1710–1755. Ph.D. dissertation, Brown University.

Mancall, P. C., Weiss, T., 1999. Was economic growth likely in Colonial British America? Journal of Economic History 59, 17–40.

McCallum, B. T., 1992. Money and prices in Colonial America: a new test of competing theories. Journal of Political Economy 100, 143–161.

McCusker, J. J., 1971. The current value of English exports, 1697 to 1800. William and Mary Quarterly 28, 607–628.

McCusker, J. J., 1978. Money and Exchange in Europe and America, 1600–1775. University of North Carolina Press, Chapel Hill, NC.

McCusker, J. J., Menard, R. R., 1985. The Economy of British America, 1607–1789. University of North Carolina Press, Chapel Hill.

Metz, W. D., 1945. Politics and Finance in Massachusetts, 1713–1741. Ph.D. dissertation, University of Wisconsin.

Michener, R., 1987. Fixed exchange rates and the quantity theory in Colonial America. Carnegie-Rochester Conference Series on Public Policy 27, 233–308.

Nettels, C. P., 1934. The Money Supply of the American Colonies Before 1720. University of Wisconsin, Madison, WI.

Newell, M. E., 1998. From Dependency to Independence: Economic Revolution in Colonial New England. Cornell University Press, Ithaca.

Newell, M. E., 2000. The birth of New England in the Atlantic economy: from its beginning to 1770. In: Temin, P. (Ed.), Engines of Enterprise: An Economic History of New England. Harvard University Press, Cambridge, MA, pp. 11–68.

O'Brien, P. K., 1985. Agriculture and the home market for English industry, 1660–1820. English Historical Review 100, 773–800.

Officer, L. H., 2003. What Was the UK GDP Then? A Data Study. Economic History Services. URL: http://www.eh.net/hmit/ukgdp/ukgdpstudy.pdf.

Osgood, H. L., 1924. The American Colonies in the Eighteenth Century, vol. 2. Columbia University Press, New York.

Perkins, E. J., 1991. Conflicting views on fiat currency: Britain and its North American Colonies in the eighteenth century. Business History 33, 8–30.

Perkins, E. J., 1994. American Public Finance and Financial Services, 1700–1815. Ohio State University Press, Columbus.

Rogoff, K., 1996. The purchasing power puzzle. Journal of Economic Literature 34, 647–668.

Shepherd, J. F., Walton, G. M., 1972. Shipping, Maritime Trade, and the Economic Development of Colonial North America. Cambridge University Press, Cambridge.

Smith, B. D., 1984. Money and inflation in Colonial Massachusetts. Federal Reserve Bank of Minneapolis Quarterly Review, 1–14.

Smith, B. D., 1985. American Colonial monetary regimes: the failure of the quantity theory and some evidence in favour of an alternate view. Canadian Journal of Economics 18, 531–565.

Smith, B. D., 1988. The relationship between money and prices: some historical evidence reconsidered. Federal Reserve Bank of Minneapolis Quarterly Review 12, 18–32.

Stark, B. P., 1984. The New London society and Connecticut politics, 1732–1740. Connecticut History 25, 1–21.

Sumner, W. G., 1876. A History of American Currency. Henry Holt, New York.

Temin, P., 2000. The Industrialization of New England, 1830–1880. In: Temin, P. (Ed.), Engines of Enterprise: An Economic History of New England. Harvard University Press, Cambridge, MA, pp. 109–152.

Weeden, W. B., 1890. Economic and Social History of New England, 1620–1789. Houghton, Mifflin, Boston.

West, R. C., 1978. Money in the Colonial American economy. Economic Inquiry 16, 1–15.

13 Afterword to Part III

Chapters 10–12 share three common properties. First, they fall within the rubric of monetary history. Second, they involve data contributions, in the form of the development of new time series for the historical experience under consideration. Third, they make use of these series to consider issues for which a quantitative approach provides insights, but for which the existing literature leaves room for improvement. Some features of the historical experiences are omitted or discussed only tangentially in the chapters. These features are given their due here – chapter by chapter, because the nature of the limitations differs in the chapters. Most of the addenda pertain to Chapter 10.

Bank Restriction Period, 1797–1821 (Chapter 10)

Historical background

The term 'Bank Restriction Period' refers to the cash suspension of the Bank of England, the most important, but by no means the only, bank in existence. Ironically, the Bank itself did make some cash payments during this period, partly for technical reasons, partly on the road to resumption of payments (see 'Analysis of resumption' below). As Chapter 10 discusses, the Bank suspension gave rise to replacement both of Britain's domestic gold standard by a paper standard and of Britain's international gold standard by a floating exchange rate.

The paper standard and floating exchange rate occurred under extraordinary circumstances. First, Britain was engaged in war – not merely the 'War of 1812' (actually occurring over 1812–14) against the United States, which had the intensity of a colonial conflict, but also 'total war' in the form of the French Revolutionary and Napoleonic Wars (1792–1815), with ultimately the very independence of Britain at stake. These were wars in which Britain participated in great coalitions, three in all, against France. Except for brief respites (March 1802–May 1803 and April 1814–February 1815), war was continuous throughout the Bank Restriction Period until Waterloo (June 1815). In adopting a freely floating exchange rate rather than a pegged rate supported by exchange control, Britain went against both its past policy under the Tudor monarchs and its future policy during the World Wars of the 20th century.

Second, the Bank Restriction Period is enveloped by the Industrial Revolution, with its concomitant economic growth and change in economic structure.

Third, the financial system exhibited a curious combination of stability and flux. On the one hand, the usury laws established a maximum interest rate on bills of exchange of five percent per year, to which certainly the Bank of England rigidly adhered (discussed in Chapter 10). On the other hand, there was a tremendous expansion of country banking early in the period, a component of the financial system highly susceptible to financial crisis. Writing more generally, Clapham (1945: 1) states: 'It is not easy to exaggerate the changes in the British banking and currency systems during the first decade of suspended cash payments at the Bank.'

Fourth, even by contemporary standards, economic statistics for the period were unusually deficient. Apart from the Bank of England, there were (and are) no reliable data about the assets and liabilities of the banking system. Also, the compilation of statistics on commodity imports and re-exports was based exclusively on fixed and outdated prices and therefore did not reflect current values. Further, a fire at the London Customs House destroyed the trade records for 1813, which forever constitutes a gap in data (see note 8 in Chapter 10).

Turning to the Bank suspension itself, historians agree that the cause of the Restriction was a gigantic loss of the Bank's gold reserves, both internationally and domestically. These historians include Angell (1926: 40–41), Ashton (1959: 134), Deane (1979: 190–93), Einzig (1970: 188), Feavearyear (1963: 177–83), Gayer, Rostow, and Schwartz (1953: 47–52), Hawtrey (1918: 53–60; 1950: 269–78), and Morgan (1943: 23–4).

The external specie drain, induced proximately by the exchange rate outside the gold-export point, took two forms. First, the balance of payments was in substantial deficit, emanating from (1) Continental war expenditure (loans and subsidies to allies, as well as disbursements for the British army), largely and reluctantly financed by the Bank at the urging of Prime Minister William Pitt, and (2) poor harvests, leading to high prices of grain and large imports. Note that these elements, albeit pertaining to a gold rather than paper standard, are consistent with the antibullionist side of the bullionist controversy.

Second, primarily because of the return of France to a metallic standard after the disastrous experience with the assignats and also because of a loss of confidence in the exchange value of the pound consequent upon unfavorable war news, guineas were melted down and exported, albeit illegally, from London to Paris by way of Hamburg. The relative importance of the two forms of the external specie drain is a matter of controversy in the literature – see Gayer, Rostow, and Schwartz (1953: 49–51). In the fall of 1796 the gold outflow apparently came to an end, but it was not followed by a gold inflow to reverse the loss in Bank reserves.

All the while, an internal specie drain proceeded, with three causes. First, war conditions involved an outward shift in aggregate demand but interrupted supply from abroad, resulting in an increase in the price level, in nominal income, and in the demand for money. By Acts of 1775 and 1777, Parliament had forbidden the issuance of bank notes of denomination below £1 and £5, respectively. Therefore the small-change component of the enhanced demand for money was obtained from the Bank's supply of guineas (and half-guineas and seven-shilling pieces), depleting its reserves.

Second, at the beginning of 1796 the Bank of England began to ration private discounts, in order to accommodate the government demand for advances (under heavy pressure from Pitt) without further expanding the Bank's note issue. However, the demand for money not satisfied by the Bank was met by note issuance of the country banks. This weakened the stability of the financial system; for the state of the reserves of the country banks was even more precarious than that of the Bank.

Third, at the end of 1796, fears of coastal raids or of an outright French invasion manifested themselves in runs on country banks. Holders of country bank notes cashed them for coin, and replenishment of the banks' cash reserves was sought ultimately in the redemption of Bank (of England) notes for guineas. In early 1797 Napoleon's successes in Austria worsened the military outlook still more, and the country banks increased their demand for gold from the Bank.

On February 25 it was reported (correctly) in the *London Gazette* that French troops had landed in Wales. Though the force was small (about 1200) and easily captured, financial panic ensued. A run on the country banks resulted in their presentation of Bank notes for redemption. Pitt feared even more that a direct run on the Bank would ensue and deplete its reserves. He summoned the Privy Council to meet on February 26 (extraordinary, on a Sunday), resulting in an Order in Council prohibiting the Bank from making cash payments until the sense of Parliament was taken on the subject. The monetary authority was clearly subordinate to the fiscal authority at this time, and, though there is no direct evidence, it is likely that Pitt envisaged suspension as a way of removing a constraint on Bank financing of government war expenditure. This interpretation is consistent with Pitt's ongoing pressure on the Bank to provide advances to the government. As Murphy (1973: 491) observes: 'With the Bank of England effectively holding a licence to print money, the government's capacity to borrow to finance war expenditure was in practice unlimited (as long as the Bank's gold holdings could support external expenditure). Under this aspect, suspension was really as much a matter of removing any check to the issue of Exchequer bills as it was of freeing the Bank from constraint.'

The inexorable path to suspension is seen not only in the Bank's gold (coin and bullion) reserves but also in the ratio of its reserves to note plus deposit liabilities, the latter measure ignored by previous writers. Only Viner (1937: 178–79) computes the Bank's reserve ratio for the Bank Restriction Period, annually but only from 1810 and without discussion. Figures used below are from semi-annual data (February 28, August 31) on Bank assets and liabilities in *Bank Charter Report* (1832, appendix no. 5: 13–25).

From reserves of £8.6 million (53 percent reserve ratio) in February 1790 in the midst of the French Revolution, there is a steady fall in both to £4.0 million (23 percent) in February 1793, a month after France declared war. Recovery ensued to £7.0 million in February (42 percent in August) 1794, but then occurred precipitous steady decline to unacceptable levels of £1.1 million (8 percent) at the beginning of Restriction.

The Bank Restriction Act of May 3, 1797 followed. It forbade the Bank from making any cash payments, with three exceptions: (1) payments for the military, (2) three-fourths repayment of a gold deposit of at least £500, and (3) redemption of notes issued on the basis of cash received after February 26, 1797. The Act did not make Bank notes legal tender. Only gold coin (and silver to a limited extent) had that status (Officer, 1996: 37). However, stipulations in the Act and subsequent legislation – Acts of 1811 ('Stanhope Act'), 1812 ('Gold Coin and Bank Note Bill') and 1814 – gave Bank notes de facto legal-tender status, and they were almost universally so treated from the first day of Restriction. [On this, see Fetter (1950) and Coppieters (1955: 36–43).] In particular, country banks redeemed their own notes in Bank notes rather than gold coin.

The original Restriction Act was to expire June 24, 1797. Successive acts renewed it until the Resumption Act ('Peel's Act') of July 2, 1819 set May 1, 1823 as the date upon which all restrictions on cash payments would cease. On the initiative of the Bank, legislation was passed in early 1821 permitting full resumption of cash payments on May 1, 1821, which in fact occurred.

Measures of depreciation

The data: exchange rate and parity

Both contemporary and later observers were intent on measuring the extent to which the pound depreciated during the Restriction Period. The purchasing power of the pound manifested itself most obviously in three markets: foreign exchange, bullion, and commodities. Considering foreign exchange, the exchange rate with Hamburg, the leading financial center of Europe during the Napoleonic Wars, is the only continuous series (apart from Lisbon – see Appendix A in Chapter 10). Amsterdam had been the financial center; but, when occupying Amsterdam in 1795, the French Revolutionary army terminated that city's pre-eminence in international finance. The same fate did not befall Hamburg, even though Napoleon's forces entered it in 1806 and again (this time looting the Bank of Hamburg) in 1813.

The relevant exchange-market instrument was the bill of exchange drawn on Hamburg and traded in London. The bill was denominated in schillings and grotes, Flemish banco. The 'Flemish' (vlamische, 'vls.') designation pertains to the Hamburg unit of account (1 pfund = 20 schillings-vls. = 240 grotes) emanating from Antwerp, the leading commercial and financial centre of Europe prior to Amsterdam. The identity of this unit of account with the British system (in fact, grotes were also called pence) was only nominal. The pound sterling always had a higher value than the Flemish pound in the foreign exchange market. Hamburg also had a second unit of account, the Lübeck ('lubs') system, shared with the nearby city of that name and involving a schilling-lubs (see below). Information on the Hamburg monetary standard is in *Bullion Report* (1810: 65, 73–5), Kelly (1811, 1821), Dickinson (1818: 106–107), Waterston (1847: 357), Shaw (1896: 387), and McCusker (1978: 62–63).

'Banco' refers to 'bank money,' transferable deposits at the Bank of Hamburg payable in silver bullion at a constant value, but rarely withdrawn except for export. Banco is distinct from 'current money' or currency, locally coined silver, used for small and common payments, and over which banco enjoyed a varying premium (agio). Also circulating were 'light money' (foreign coin) and 'specie' (full-weight silver coins, minted formerly, with an agio above banco). The two alternative units of account applied to each of the four monies, and agios varied; but bills of exchange were expressed and payable only in banco.

The schilling–pound exchange rate was given by the bill's schilling face-value/pound market-price ratio. The pound payment and schilling receipt were not concurrent. Foreign bills were transacted in London only on the 'post days,' Tuesday and Friday, when the mails left for abroad; and bills were payable on the subsequent post day – see Clare (1893: 38) and Einzig (1970: 177). Therefore the bill purchaser (payer of foreign exchange abroad), while receiving the bill immediately, did not make pound payment in London until 3–4 days afterward.

The bill seller was the drawer of the bill and issued the bill to the drawee abroad. The payee presented the bill to the drawee, receiving the foreign exchange. Availability of foreign exchange to the payee abroad (who was mailed the bill by the payer) depended on three elements: (1) the usance (unit for measuring the length of time between creation of the bill and its maturity), (2) number of usances applicable, and (3) days of grace (additional days allowed for payment beyond the bill's maturity, specified by the law of the drawee city). For bills on Hamburg, the usance was one month after the date of the bill, 2½ usances applied, and there were 12 calendar days of grace – see Kelly (1811: 204, 276; 1821: 178) and McCulloch (1851: 64–65). Therefore schillings were certain to be available approximately 2.9 months after purchase of the bill and 2.8 months after payment in pounds.

This is the basis of exchange-rate data, expressed in schillings and grotes per pound, and taken from weekly tabulations in *Resumption Report* (1819: 335–54) for 1796–1818 and *Bank Charter Report* (1832: 98–100) for

1819–1821. The data are converted to schillings and averaged quarterly 1796–1821 to obtain:

ER = exchange rate, schillings per pound

The series ER is presented in Table 13.3 below. To obtain the mint parity corresponding to this exchange rate, the official valuation of the metallic standard in each center must first be obtained. In 1663, by Royal Order, it was established that 44½ guineas were to be coined from one pound Troy (5760 grains) of standard (11/12th) fineness gold, implying a guinea weight of 129 39/89 grains. In 1717 a Royal Proclamation set the price of the guinea at 21 shillings, implying that a pound sterling is the value of (129 39/89)·(11/12)·(20/21) = 113 1/623 grains of pure gold. The Coinage Act of June 22, 1816 stated, in effect, that future issues of gold coin should have the fineness and value-to-weight ratio of the guinea; and the sovereign, valued at 20 shillings and weighing 123.274 grains of standard gold according to Royal Proclamation, was so coined beginning July 1817. [The Proclamation truncated the weight of the sovereign; its accurate weight consistent with the guinea was 123 171/623 grains. On all this, see Officer (1996: 35–36).]

While a gold-coin standard was in suspension in Britain, a silver-bullion standard existed at Hamburg. The Bank of Hamburg transacted in silver bars (47/48th fineness) at prices expressed in the Lübeck unit of account (1 mark = 16 schillings-lubs = 192 pfennigs). Since 1790, the Bank's buying (selling) price was 27 marks, 10 (12) schillings-lubs banco, per Cologne mark (equal to 3608 grains Troy) of pure silver. For computation of parity, the applicable price is that at which silver is issued. The equivalence 1 mark = 2⅔ schillings-vls. enables conversion between the Flemish and Lübeck units of account, whence 27 marks, 12 schillings-lubs translates to 74 schillings-vls. Then one schilling-vls. banco is the value of 3608/74 = 48 28/37 grains of pure silver.

The ratio (113 1/623)/(48 28/37) = 2.317660+ has dimension (number of schillings per pound)/(number of grains of silver per grain of gold). For a meaningful mint parity, it must be multiplied by the market price of gold in terms of silver (number of kilograms of silver per kilogram of gold). The best data are in Soetbeer (1879: 130), though available only annually. The resulting product yields the annual series:

PAR = mint parity, schillings per pound

This measure of parity, while accurate to more significant digits, is identical conceptually to that computed by Silberling (1919: 287–88) and Marcuzzo and Rosselli (1990: 115–19). However, these authors err in calculating the percentage deviation of the exchange rate from PAR; because, in contrast to the exchange rate, PAR implies a simultaneous exchange of schillings for pounds. As noted by Ricardo (1811; 1951: 168) and Hawtrey (1918: 54), measured parity must be increased to allow for interest loss suffered by the bill purchaser. They stipulate a waiting period of 2.5 months rather than the more precise figure of 2.8 months and do not compute a time series. Corrected mint parity, on the same basis as the exchange rate, is given by:

$$PAR^* = [1 + (2.8/1200) \cdot INT] \cdot PAR$$

where INT = interest rate, percent per year.

There is a large, mainly 20th-century, literature on whether the interest rate to convert a time to a sight bill should be that in the drawer or drawee city, and in principle that literature applies to INT. For references to the literature, see Collins (1986: 514) and Officer (1996: 69; 297, ns. 29–30). Contemporary observers of the Restriction Period were unanimous that the London rate applied to this, or similar situations; and INT is derived on that basis. Walter Boyd, testifying in *Report of 1797* (1810: 65), makes a correspondence between bills drawn in London on Hamburg (Hamburg on London) and the London (Hamburg) interest rate. A Continental merchant (identified as John Parish, Jr., by Sraffa, 1951: 427–34) examined in *Bullion Report* (1810: 74) states that the interest rate implicit in the differential between the exchange rates at London and at Hamburg is five percent (known to be the London rate). Ricardo (1811; 1951: 168) notes that, because bills are at 2½ months and par implies simultaneous exchange, for proper computation of the deviation of the (London on Hamburg) exchange rate from par, an allowance for interest of 'about 1 per cent' must be made, which results from a five percent interest rate. Therefore INT is obtained via the London, not Hamburg, interest rate.

Because bills are the exchange-market instrument, the *market* interest rate is the logical concept for INT. No data on that rate exist for the Restriction Period, but information to construct INT is available. The usury laws, setting a five percent limit on annual interest, applied to bills of exchange throughout the Restriction Period. The discount rate of the Bank of England was also five percent. For both these reasons, the market discount rate (for good bills, those eligible for Bank discounting) did not exceed five percent during the Restriction. Apart from the avoidance techniques of bill brokers (charging commission) and private bankers (requiring minimum balances), the usury laws were generally obeyed. The penalty for illegal interest was high, triple the capital value of the transaction. Beginning July 1817 the market rate fell below five percent, remaining there for about a year. A rate of 4½ percent is taken for INT for 3Q 1817–2Q 1818, 5 percent otherwise. (For interest rate references, see Chapter 10, note 5.) Corrected mint parity, PAR*, an annual series except for 1817–1818, is presented in Table 13.1.

Retrospective measures of depreciation

There are two categories of measures of depreciation of the pound. The first is retrospective, with norm the value in a base period, taken here to be the average over the four quarters of 1796. The price level falls into the

Table 13.1 Corrected mint parity, 1796–1821 (schillings per pound)

Year	*Parity*	*Year*	*Parity*	*Year–Quarters*	*Parity*
1796	36.69	1806	36.39	1816	35.83
1797	36.13	1807	36.18	1817–1Q, 2Q	35.43
1798	36.55	1808	37.70	1817–3Q, 4Q	35.39
1799	36.91	1809	37.42	1818–1Q, 2Q	35.95
1800	36.76	1810	36.98	1818–3Q, 4Q	35.99
1801	36.25	1811	36.41	1819	35.94
1802	35.78	1812	37.77	1820	36.62
1803	36.13	1813	38.10	1821	37.40
1804	36.13	1814	35.26		
1805	37.02	1815	35.78		

Source: see text.

retrospective category, and the series used is the monthly index number of domestic and imported commodities developed by Gayer, Rostow, and Schwartz (1953: 468). The index number is rebased and averaged quarterly to yield

PL = price level, 1796 = 100

The PL series is shown in Table 13.3 below. To conform to the price level, the exchange rate ER is inverted (so that a higher value represents greater depreciation of the pound), and based on 1796:

ERINV = exchange-rate inverse, pounds per schilling, 1796 = 100

The variables PL and ERINV are graphed in Figure 13.1. Except for small intervals, the price level exhibits greater depreciation of the pound than does the exchange rate, and is a more-volatile series. Maximum PL, of 151, occurs in 1Q 1801, while maximum ERINV, at 142, occurs in 2Q 1811. In contrast, Hawtrey (1918: 62–5; 1950: 279–92) and Vilar (1991: 311–12), the modern authors who systematically study the pound depreciation of the Restriction Period, observe a relatively lower, and only a local, maximum in 1801. For them, the price level peaks during 1809–1814. Their work suffers from use of only annual data, incorporation of earlier and inferior price indices, and failure to examine the index number of the pure exchange rate.

Prospective measures of depreciation

The second category of depreciation measures is prospective, with the value to which the measure is expected to return serving as norm. Corrected parity, PAR*, is clearly the prospective norm for the exchange rate. Working with inverses, the exchange-rate measure becomes 100·(PAR*/ER), denoted as

ERINVP = exchange-rate inverse, pounds per schilling, parity = 100

While a prospective norm value is not known for the price level, it does exist for standard (11/12th-fineness) bar gold: the mint price, £3 17s. 10½d., or 934½d., per ounce. With the pound sterling having value (12/11)·(113 1/623) = 123 171/623 grains of standard gold, and with 480 grains = 1 oz., the mint price of standard gold is 480/(123 171/623) = £3.89375, or £3 17s. 10½d., per ounce. The price of standard gold in the London market is available weekly (same sources as for ER). It is converted to pence and averaged quarterly, yielding a price series for standard gold, PG; but there are many missing observations (1Q 1796, 1Q 1800–1Q 1804, 1Q 1806–3Q 1810, 2Q–3Q 1811, 3Q 1812), for lack of market transactions. The depreciation measure, obtained as 100·(*PG*/934.5), is:

PGP = price of standard gold, mint price = 100

The variables ERINVP and PGP are graphed in Figure 13.2. Remarkable is how close the measures are from 4Q 1812 onward, when the gold price is continuously observed. Using annual data, Hawtrey (1918, 1950) found the maximum depreciation of the pound measured by the exchange rate and the price of gold to be 1811 and 1813, respectively. Quarterly data here confirm

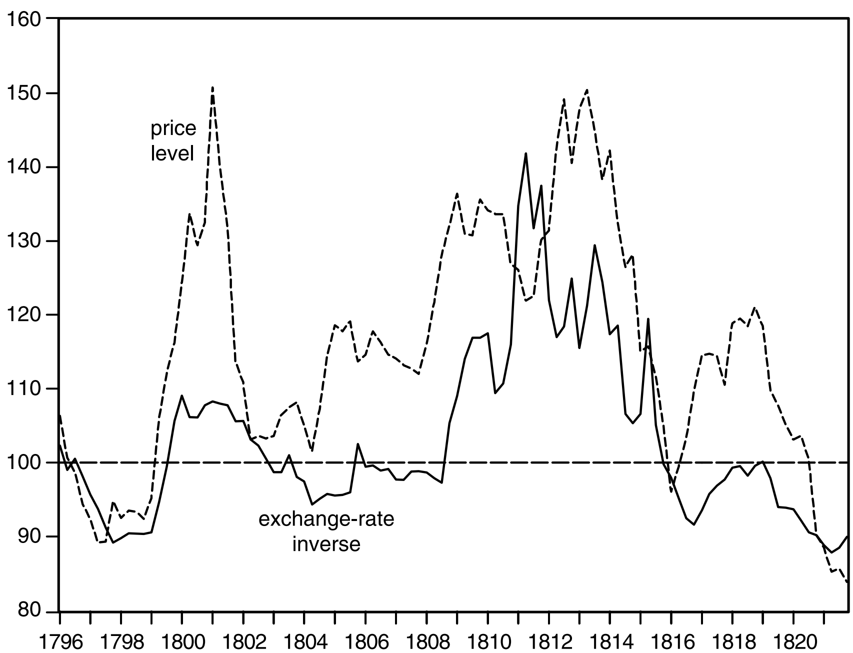

Figure 13.1 Measures of depreciation of pound, 1796–1821 (1796 = 100).

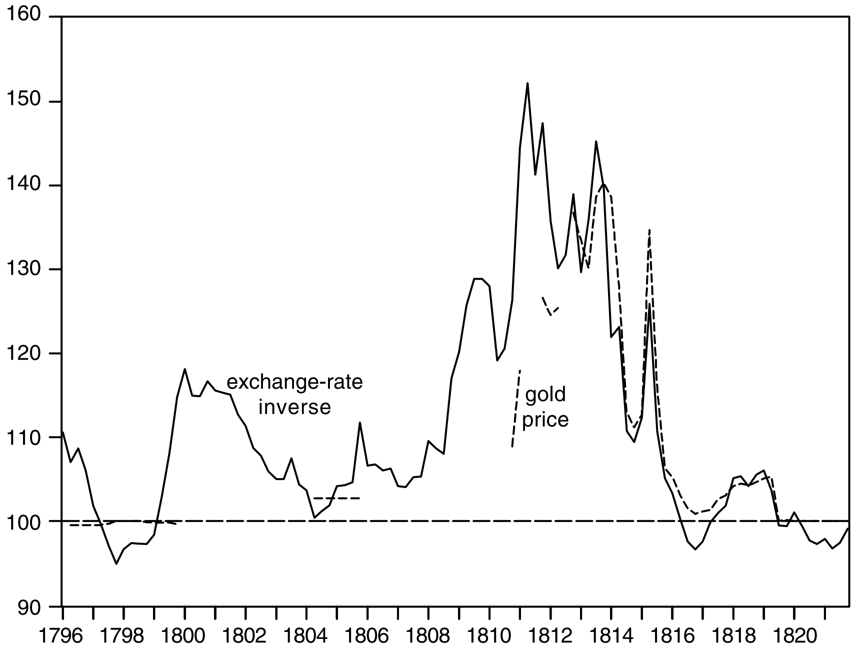

Figure 13.2 Measures of depreciation of pound, 1796–1821 (par = 100).

the result; maximum ERINVP (152) and PGP (140) occur in 2Q 1811 and 4Q 1813.

Whether the norm value of the exchange rate is the 1796 level or mint parity makes a substantial difference in whether or not the pound is depreciated. Only 44 of the 96 quarters 2Q 1797–1Q 1821 exhibit depreciation under the former criterion (ERINV > 100), but 78 under the latter (ERINVP > 100). Which measure of the exchange rate is appropriate depends on the issue at hand. For exchange rate determination, the pure exchange rate (ER or ERINV) is the relevant dependent variable. For analysis of resumption, the deviation of the exchange rate from parity (ERINVP) is the pertinent measure.

Analyses of resumption

Traditional analysis

Existing studies of the Bank's path to resumption center primarily on the amount of Bank reserves, secondarily on the price of gold over the mint price and the exchange rate over parity. See Acworth (1925: 69–114), Canaan (1925: xxix–xxxiv), Viner (1937: 171–85), Clapham (1945: 42–43, 62–75), Gayer, Rostow, and Schwartz (1953: 52–53, 133–34, 163–65), and

Feavearyear (1963: 190–91, 214–24).The analysis of these authors is improved here in three ways: (1) using quarterly rather than annual data, (2) considering reserves not only in absolute amount but also as the ratio to note and deposit liabilities, and (3) correcting parity for non-contemporaneous exchange of domestic for foreign currency.

In November 1816 the Bank offered to redeem in gold its £1 and £2 notes dated prior to 1812. This positive move toward resumption differed from the Bank's practice, ongoing since 1799, of paying cash as change for its £5 notes. Also, in 1798 the Bank had redeemed in cash its first issue of small notes, before the type was changed. The motivation of these policies was convenience. In contrast, the 1816 policy was a move toward resumption. At the end of August 1816, Bank reserves (£7.6 million, 20 percent ratio) were as high as they had been since early 1808. In 3Q 1816 the gold price was the lowest recorded since 1799 and only slightly above par; the exchange rate was below parity (ERINVP < 100) for the first time since 1Q 1799. It was a propitious time to move toward resumption, and there was hardly any response to the Bank's offer – a good sign.

In February 1817 the Bank's reserves were at an all-time high, £9.7 million, and the reserve ratio had climbed to 25 percent; in April the Bank extended its offer to incorporate notes dated prior to 1816. Again there was little demand for redemption; the Bank's reserves increased to £11.7 million (30 percent ratio) in August, and in September the offer was extended to notes dated through 1816. By this time, both the exchange rate and gold price were above parity and increasing; notes were cashed in substantial amount, and the Bank's reserves fell to £4.2 million (13 percent ratio) in February 1819, shortly after resumption committees had been appointed in Parliament.

The result was, first, the Act of April 6, 1819, forbidding Bank payment in gold until further action, and, second, the Resumption Act of July 2, 1819, which legislated a graduated return to the gold standard. The mechanism was a transitional bullion standard, with the Bank redeeming its notes in gold bars at decreasing prices in stages, reaching the mint price on May 1, 1821, and switching to payment in coin on May 1, 1823. There was no demand for bars, the gold price and exchange rate fell precipitously almost to parity or below in 3Q 1819, Bank reserves briefly decreased but then rose to a record £11.9 million (40 percent ratio) in February 1821, and redemption in coin occurred on May 1, 1821.

Gold-point analysis

When the exchange rate (ER) is between the gold points, there is good evidence that a return to the gold standard is sustainable. Because of (1) the high, varying, and largely unknown insurance component of the gold points under wartime conditions, and (2) the difficulty of computing gold points when melting and export of domestic coin were ongoing though illegal (true until the Resumption Act), there are only scattered estimates of gold points during the Restriction Period [see *Report of 1797* (1810: 56, 64) and *Bullion Report* (1810: 11–12, 76)]. The problems are obviated here by constructing hypothetical gold points, under the counterfactual assumptions that the gold standard is in existence and that there is peacetime throughout 1797–1821.

The prospective gold exporter in London has two alternative methods of transferring funds to Hamburg: (1) purchase a bill of exchange in London and mail it to Hamburg, (2) purchase gold in London, ship it to Hamburg, exchange it for silver, and obtain bank money for the silver. Similarly, the prospective gold importer in London has two ways of transferring funds from Hamburg: (1) sell a bill of exchange, (2) redeem bank money for silver in Hamburg, exchange the silver for gold, ship the gold to London, and sell the gold. The gold point is the cost of (2) minus the cost of (1).

Derivation of the gold points, as a percentage of the amount transferred, or, equivalently, as a percent of parity, is shown in Table 13.2. Letting {GXP, GXPER} ({GMP, GMPER}) denote the gold export (import) point in {schillings per pound, percent of parity},

$$GXP = (1 - GXPER/100) \cdot PAR^*$$

$$GMP = (1 + GMPER/100) \cdot PAR^*$$

Table 13.2 London–Hamburg 'peacetime' hypothetical gold points, 1797–1821, cost components (percent)

Item	*Export point*	*Import point*
I. Direct cost		
A. Transportation		
1. Sea portion: freight	0.2500	0.2500
2. Land transport		
a. Hamburg to/from Cuxhaven	1.0000	1.0000
b. London to/from Yarmouth	0.7500	0.7500
B. Insurance	0.5250	0.5250
C. Commission at Hamburg	0.3333	0.3333
D. Melting in London	0.2140	0.2140
II. Interest cost		
3Q 1817–2Q 1818	–0.7500	0.1875
All other quarters	–0.8333	0.2083
III. Total cost		
3Q 1817–2Q 1818	2.3223	3.2598
All other quarters	2.2390	3.2807

Source:
I.A, I.B., I.C. – *Report of 1797*, Minutes of Evidence (1810: 56). Stipulated costs are for gold import, but charges for gold export are stated to be 'much alike.'
I.D. – *Bullion Report*, Minutes of Evidence (1810: 47). Projected from expense of melting light guineas into bars, stated to be two pence per standard ounce, corresponding to a charge of 200/934.5 percent at the mint price of £3 17s. 10½d. II. – see text. III. – sum of direct cost (I) and interest cost (II).

Among the assumptions in the derivation of GXPER and GMPER are: (1) gold is bought or sold in the London market at the mint price; (2) gold from one center has to be melted and recast into suitable bars for sale in the other center, and melting takes place in London; (3) brokerage for the purchase or sale of bills of exchange, an unknown rate, equals brokerage for the purchase or sale of gold (1/8th percent – *Bullion Report* 1810: 37), thus canceling out.

The transaction between gold and silver at Hamburg is incorporated in the definition of PAR*, with commission listed in Table 13.2. Interest cost is negative for gold export, a two-month gain. Gold may be converted into cash or bank money 'full two months before . . . bills.' – *Report of 1797* (1810: 63). The background is a 2½-month bill pure and simple. This suggests a half-month loss for gold import. Interest cost is computed as k·INT, where k = –2/12 (+0.5/12) for gold export (import). The gold points, GXP and GMP, along with the exchange rate (ER), are plotted in Figure 13.3. There are two reasons why the gold points are variable over time. First, INT falls below five percent for the four-quarter period 3Q 1817–2Q 1818. Second, PAR* changes annually, or semi-annually 1817–1818 (but not quarterly, which explains the jagged pattern of the gold points).

The position of the exchange rate in relation to the gold points shows the need for Restriction even under hypothetical continuous peacetime conditions. Of the 96 quarters of Restriction, only 3 involve a gold-import point violation but 71 a gold-export point violation. With final peace following Napoleon's second abdication in June 1815, the gold points pertain to actual

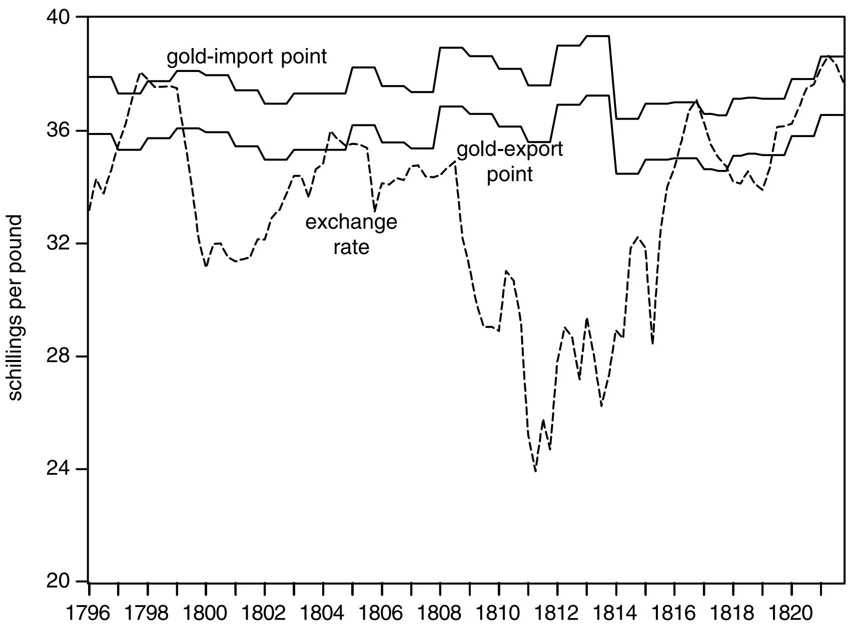

Figure 13.3 Exchange rates and 'peacetime' hypothetical gold points, 1796–1821.

rather than hypothetical peacetime conditions. In 2Q 1816 the gold-export point ceases to be violated for the first time since 1804, and the Bank's limited-redemption policy is justified. With the exchange rate below the gold export point 1Q 1818–2Q 1819, re-imposition of full suspension in April 1819 is understandable. Thereafter the exchange rate stays within the gold-point spread, leading to early resumption.

Tabulation of variables

Unlike Chapters 11 and 12, Chapter 10 does not exhibit time series of important variables. To correct this omission, Table 13.3 presents the data – exchange rate, price level, Bank of England notes, price of wheat, and external military expenditure – that are the subjects of empirical analysis in chapter 10.

Table 13.3 The data, 1797–1821

Year: quarters	*Exchange rate*[a] *(ER) (schillings per £)*		*Price level*[b] *(PL) (1796 = 100)*		*Bank of England notes (BN) (£m)*		*Price of wheat (PW) (1796 = 100)*		*External military expenditure (ME) (£m)*	
97:1–2	35.49	36.23	92.44	89.14		11.75		64.63		0.52
97:3–4	37.21	38.08	89.29	94.80	11.00	11.39	70.12	73.17	0.49	0.38
98:1–2	37.81	37.54	92.50	93.48	12.88	13.02	65.24	67.07	0.09	0.05
98:3–4	37.56	37.58	93.31	92.39	12.15	12.24	65.85	62.80	0.06	0.12
99:1–2	37.51	35.85	95.20	105.70	12.99	13.76	64.63	76.83	0.50	0.67
99:3–4	34.13	32.14	112.08	116.21	13.36	13.78	93.29	115.24	0.78	0.83
00:1–2	31.13	31.98	124.33	133.77	15.02	14.96	131.71	151.83	0.88	0.93
00:3–4	32.00	31.51	129.41	132.48	15.13	15.49	146.34	155.49	0.91	0.82
01:1–2	31.36	31.44	150.82	139.89	16.44	15.82	190.24	177.44	0.68	0.59
01:3–4	31.50	32.15	131.74	113.77	15.29	15.68	150.00	96.95	0.51	0.42
02:1–2	32.13	32.92	110.87	103.15	15.59	16.79	96.95	87.20	0.30	0.22
02:3–4	33.19	33.77	103.58	103.24	16.96	17.38	87.80	77.44	0.17	0.15
03:1–2	34.40	34.40	103.61	106.45	15.66	16.23	73.17	76.22	0.12	0.09
03:3–4	33.61	34.63	107.46	108.20	16.78	17.27	73.17	70.12	0.08	0.09
04:1–2	34.85	35.99	104.96	101.51	17.63	17.61	65.85	67.07	0.07	0.07
04:3–4	35.70	35.47	107.20	114.29	17.13	17.25	76.83	100.61	0.12	0.20
05:1–2	35.53	35.50	118.59	117.76	17.62	16.93	116.46	116.46	0.51	0.65
05:3–4	35.38	33.12	119.11	113.68	16.49	16.49	120.12	101.83	0.71	0.66
06:1–2	34.14	34.08	114.57	117.76	16.85	17.02	96.95	106.10	0.48	0.45
06:3–4	34.32	34.24	116.29	114.57	16.67	16.61	106.10	100.61	0.47	0.53
07:1–2	34.74	34.76	114.08	113.17	16.66	16.79	99.39	98.17	0.58	0.64
07:3–4	34.38	34.35	112.68	112.02	16.97	16.40	95.12	87.80	0.80	1.06
08:1–2	34.41	34.69	115.92	121.69	16.65	17.18	89.63	96.95	1.60	1.91
08:3–4	34.90	32.22	128.09	132.05	17.17	17.42	106.71	116.46	2.11	2.19
09:1–2	31.13	29.77	136.39	130.99	17.84	18.53	122.56	118.29	2.15	2.23
09:3–4	29.04	29.04	130.76	135.64	19.34	19.94	122.56	134.15	2.29	2.34
10:1–2	28.89	31.03	134.18	133.66	20.44	21.35	131.71	143.90	2.26	2.31
10:3–4	30.67	29.27	133.66	126.91	24.17	24.21	146.95	129.27	2.46	2.73
11:1–2	25.19	23.92	126.05	121.92	23.33	23.61	122.56	114.02	3.18	3.46
11:3–4	25.77	24.69	122.55	130.16	23.28	22.91	117.68	134.76	3.70	3.91
12:1–2	27.83	29.03	131.42	142.64	23.32	22.92	139.63	169.51	3.92	4.11
12:3–4	28.67	27.18	149.19	140.55	23.46	23.32	187.20	153.66	4.46	4.95
13:1–2	29.38	28.00	147.89	150.45	23.93	23.95	156.10	153.66	5.88	6.41

13:3–4	26.23	27.29	145.02	138.31	23.96	24.26	142.07	110.37	6.68	6.68
14:1–2	28.92	28.63	142.24	132.48	25.16	25.86	100.61	92.68	6.76	6.76
14:3–4	31.85	32.22	126.45	128.18	28.64	27.97	95.12	94.51	6.43	5.77
15:1–2	31.85	28.42	115.06	115.72	27.30	27.01	96.34	90.85	4.94	4.30
15:3–4	32.32	34.03	111.62	105.13	27.17	26.07	86.59	73.17	3.59	2.81
16:1–2	34.67	35.70	96.06	99.50	26.57	26.36	70.12	90.85	1.49	0.91
16:3–4	36.72	37.08	103.26	109.69	27.23	26.13	104.88	126.83	0.55	0.40
17:1–2	36.29	35.47	114.49	114.72	27.14	27.54	132.93	139.02	0.07	0.03
17:3–4	35.04	34.74	114.37	110.55	29.50	28.92	115.85	104.27	0.01	0.02
18:1–2	34.19	34.12	118.85	119.51	28.42	27.49	109.76	112.20	0.13	0.16
18:3–4	34.56	34.10	118.45	121.03	26.89	26.01	107.93	106.10	0.18	0.19
19:1–2	33.90	34.69	118.50	109.81	25.79	25.39	103.05	93.90	0.19	0.20
19:3–4	36.13	36.16	107.77	105.16	25.48	23.91	95.73	86.59	0.20	0.18
20:1–2	36.24	36.83	103.12	103.64	24.13	23.76	85.37	90.24	0.15	0.13
20:3–4	37.49	37.65	100.60	90.23	24.46	23.28	90.85	73.78	0.11	0.09
21:1–2	38.21	38.65	88.51	85.21	24.16		69.51		0.06	

a *1796:1–4* 33.19 34.29 33.77 34.58
1821:3–4 38.38 37.74
b *1796:1–4* 106.31 100.65 98.70 94.34
1821:3–4 85.61 83.86
Source: see Chapter 10, Appendix A.

U.S. specie standard, 1792–1932 (Chapter 11)

The existing literature on the economic performance of the gold standard, in Section VIII of Chapter 11, does not receive sufficient recognition. A brief rectification is made here. To assess the empirical record of the gold standard, the usual measures relate to price and per capita real income. Performance of a monetary standard is deemed higher: (1) the lower is average price inflation or deflation over time (the closer absolute inflation is to zero), (2) the lower is the variability of inflation, (3) the higher is the growth of real per capita income, and (4) the lower is the variability of income. Typical are the studies of Bordo (1993) and Bordo and Schwartz (1999), who jointly consider the four core countries (United Kingdom, United States, Germany, France) as well as Canada, Italy, and Japan, and compare performances of the classical gold standard (1881–1913), interwar period (1919–1938), Bretton Woods (1946–1970) and floating exchange rate (1974–1989 or 1974–1995). For all countries except Japan, the classical gold standard performs best according to criterion (1); but for no country does the classical gold standard excel under the other criteria.

Chapter 11 (Section VIII) examines the United States while on the contingent specie [silver or gold] standard (1792–1932), and compares the economic performance of the classical gold standard (1879–1913) with periods of central banking (1792–1810, 1817–1838, 1914–1932), the Independent Treasury system (1847–1861) the greenback period (1862–1878), and other periods (1811–1816, 1839–1846). The classical gold standard performs best according to criteria (1) and (2), and is second, trivially behind the greenback period, according to criterion (3). Applying aggregate (rather than per capita) real income as the measure for criterion (4), Chapter 11 finds that the classical gold standard is decidedly superior to the greenback period, although behind the 1811–1816 and 1817–1838 periods. In sum, at least for the United States, the classical gold standard undeniably has the best economic performance to the end of the gold standard (1933) – but the comparison is only over time for the given country (the United States), rather than a combined intercountry-intertemporal investigation, which is the hallmark of the literature.

It should be commented, however, that intercountry measurements of the economic performance of the interwar gold standard are not particularly meaningful, because the four core countries together were on the gold standard for less than four years (1928–1931) and because of the impingement of the Great Depression. Nevertheless, it is fair to say that the combination of the gold standard, which restricted monetary and fiscal policy, and the Great Depression, which cried out for expansionary policies, worsened economic performance.

Quantity theory in New England, 1703–1749 (Chapter 12)

Chapter 12 develops time-series of the components of the 'equation of exchange.' More can be stated about the measure of each component, via reference to the literature. The components are the price level, money stock, real income, and velocity.

Price level

Three types of annual price variables for New England for the 1703–1749 period exist. First are direct measures of commodity prices in the region, based on the only available price series: Boston wholesale prices of wheat and molasses, in Cole (1938, p. 117), and used not only by West (1978) but also categorically by Smith (1985: 542–44). As argued in Chapter 12 (Section 3), focusing on only one or two commodities can hardly be a good measure of the general price level. A similar judgment is made by West (1978: 3–5) and Brock (1992: 5–6).

Second, there is the price of silver, which was often used by New Englanders to measure the depreciation of their paper currency (Weiss, 1974: 588). A disadvantage is that silver played several roles: an actual or potential medium of exchange, a store of value, and a commodity itself. Interpretation of the changing price of silver as purely a reflection of inflation is by no means obvious. Brock (1992: 10) claims that empirical evidence supports 'the price of silver . . . as a rather effective measure of inflation in New England;' but, according to the related and more-specific data in Weeden (1890: 473, 677–78), this is not so.

Third, the £-Massachusetts /£-sterling market exchange rate, often used by modern scholars to represent the New England price level, has the advantage of being based on purchasing power parity (PPP) – a monetarist theory consistent with the quantity theory of money. Furthermore, general price levels are implicit in the theory and application; so dependence on one or two commodities vanishes. These advantages are overwhelming, in this author's view. The problem is that the PPP justification of the measure involves the assumption of constancy of the British price level for half a century! To their credit, McCallum (1992: 149) and Sumner (1993: 145) make this assumption explicit.

The price-level measure adopted in Chapter 12 involves applying PPP explicitly, using both the exchange rate and a newly assembled index of the British price level. While an improvement over existing indicators of the price level, this PPP-based price variable has the limitation that transportation costs were high and variable, as mentioned in Chapter 12 (Section 3). So PPP could be satisfied only approximately. Therefore the new price variable (P) is best viewed as a corrected exchange-rate proxy of the price level.

Brock's (1975) preferred price variable for New England – the product of the Boston/Philadelphia wheat price ratio and the sterling exchange rate – is in the category of the first type of price variable (with the limitations thereof), because the exchange rate pertains to Philadelphia (rather than Boston) on London. See Brock (1975: 594–95; 1992: 10).

Figure 13.4 traces percentage deviations of the silver price (from Brock 1992, table 2) and exchange rate from the new price series. Allowing for changes in the British price level makes a difference: 35 of the exchange-rate divergences are negative, with one at 20 percent and four others exceeding 10 percent. In contrast, the silver price is generally – and substantially – above

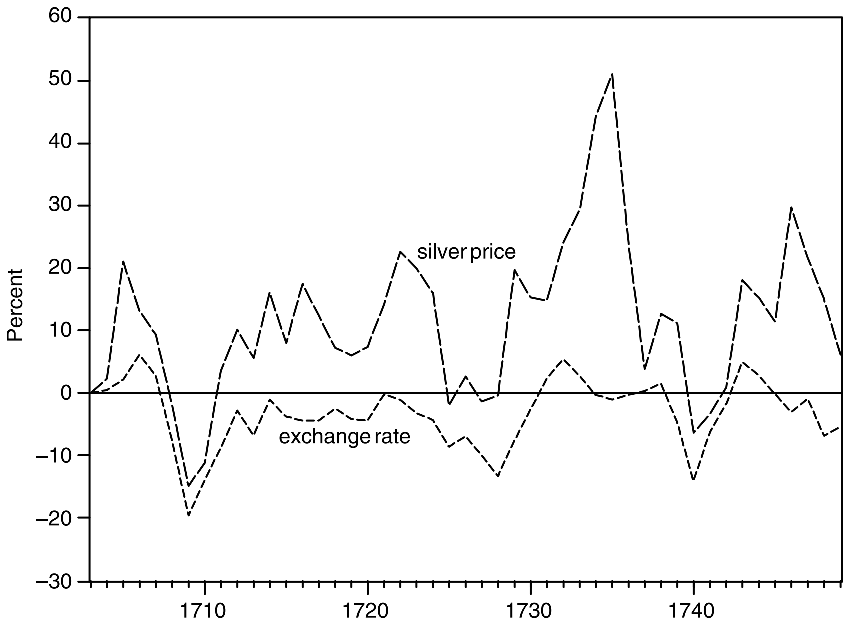

Figure 13.4 Deviations from price level, 1703–1749.

the price series: of 38 positive divergences, the maximum is 51 percent, with 25 others exceeding 10 percent.

Components of money stock

Public bills of credit and private notes

Little need be added to the discussion in Chapter 12 (Section 3). As mentioned there, Brock's (1975) public bills series has a timing inconsistency, corrected in Chapter 12. Regarding valuation, Brock (followed by McCallum) measures Boston Merchants' notes at face value, even though they went to a premium over public bills – again the warranted correction is made in Chapter 12.

Other private notes are ignored by these authors, with Brock alone providing justification. Brock (1975: 52–53) argues that 'since the notes of the [New London] society were replaced by those of the colony, one does not need to take separate account of them in determining the amount of bills in circulation,' and refers to 'the apparently abortive' note issuance of New Hampshire merchants. He also claims that: 'The Land and Silver Banks of 1740 need not engage our attention as they soon were suppressed by Act of Parliament and had little economic effect' (Brock, 1992: 10). These arguments simply reflect the fact that the private notes under consideration had only a temporary circulation – a circulation that merits recognition in a comprehensive money stock series, as developed in Chapter 12.

Specie in circulation

As noted in Chapter 12 (Section 3), there exists no directly measured series of the specie stock in New England (or any colony). However, several techniques to estimate specie in circulation have been used by scholars. Letwin (1981: 465–68) applies the classical quantity theory to Pennsylvania and all colonies jointly, assuming fixed annual velocity (at 10) and per capita income (at £20), to compute the money stock, then subtracting paper currency to obtain gold. The necessary assumptions regarding velocity and income, implicit in this method, are so strong and arbitrary as to call into question the validity of using the resulting series to perform analysis.

Grubb (2004: 334–35) estimates the Pennsylvania ratio of specie to currency in circulation annually as the average ratio of specie to paper currency offered in published awards for runaways. Specie in circulation is obtained as the product of that ratio and the known outstanding paper currency. Grubb's approach is ingenious, but in practice the high variability in the computed ratio from year to year suggests that the resulting specie series could be unreliable.

The technique adopted in Chapter 12 is a refinement of that applied graphically by Davis (1900: frontispiece) and Brock (1975: 30e, 30i [reproducing

Davis]): accept contemporary specific-year estimates of silver in circulation, and interpolate for the intervening years. A contemporary estimate of circulation in an intermediate year is used by these authors. However, the intermediate figure is both an overestimate and is graphed incorrectly. In any event, and incomprehensibly, the authors' specie estimates go nowhere: neither Davis nor Brock combines silver with currency in circulation to generate a money stock series.

Bills of exchange

Bills of exchange, negotiable liabilities issued by a 'drawer' on the basis of available funds in London, are not mentioned in Chapter 12 but deserve recognition. Bills of exchange had some circulation in New England. Unlike bills of credit, bills of exchange (1) were denominated in pounds sterling rather than New England pounds, (2) had a definite, and relatively short, maturity date, (3) were payable in London rather than in the colony, and (4) had a fluctuating value in New England currency, depending on time to maturity and quality of the bill (identity of the drawer). An advantage of bills of exchange is that, along only with specie, they could be used to settle balance of payments deficits with England.

However, there are two good reasons why the bills had only a limited medium of exchange function. First, their value, although in principle flexible, was generally high, typically £100 (McCusker 1976: 102). With the Spanish dollar worth 4s. 6d. sterling, the typical bill of exchange was therefore a multiple of 444 4/9th the Spanish dollar in value – far too high to serve as a medium of exchange in ordinary transactions. Second, bills could be dishonoured and there were expenses in recovery, which discouraged people from purchasing bills of distant origin. Further, 'equally liquid large denomination liabilities are not included in modern money supply measures' (Smith, 1987: 9), and even contemporary usage did not count bills of exchange as money. On balance, in measuring the New England money stock, bills of exchange are legitimately excluded.

Book credit

The dismissal of book credit from the money stock, in Chapter 12 (note 17) is too glib. Book credit, meaning credit extended by merchants and transferable 'on their books,' is seen by some scholars – Baxter (1945), Ernst (1973), West (1978), Michener (1987), and apparently Flynn (2001) – as the principal component of the New England money stock. A discussion of book credit from the standpoint of these authors is in order. These scholars view merchants as playing the role of bankers, with book credit analogous to demand deposits. Whether book credit in fact had the importance so ascribed to it is unknowable, because (just as for bills of exchange) data on outstanding book credit are lacking. Some historians, such as Hanson

(1979) and Smith (1985, 1987), are skeptical. Hanson points out that book credit had deficiencies as a medium of exchange: uncertain creditworthiness of customers, costs of collection, etc. Smith notes that trade credit and credit-card balances, analogous to book credit, are excluded from the modern money supply.

Also, Smith properly observes that contemporary usage counted only specie and paper currency as money. However, it is possible that this practice simply reflected the fact that specie and paper currency were the means of settling book credit balances (the primary money). In that case, the monetary aggregate 'specie plus paper currency' may be interpreted as the monetary *base* rather than the money stock.

Income

As stated in Chapter 12 (Section 4), directly measured data on colonial income do not exist. Therefore scholars have had recourse to indirect techniques. One method is to assume income is proportional to exports and/or imports (for which data exist), perhaps adjusted for other information. An example is Egnal (1998), who applies the technique to three regions of the 13 colonies. A similar approach involves probate-inventory wealth in place of international trade (for references, see Grubb 2004: 351–52). The problem with this technique is that the income/trade (or income/probate-inventory-wealth) ratio is either assumed constant, which is overly simplistic; or adjusted subjectively, which is arbitrary. It is not surprising that users of the method present their estimates as growth rates rather than time series.

The second method, used by Grubb (2004: 352–56) for Pennsylvania, is to compute (per capita) real income from the equation of exchange, using (1) data on the money stock, price level, and population, and (2) an assigned value for velocity. The weakness of this technique is that velocity is unknown, so its value must be assumed. To Grubb's credit, he offers alternative values of velocity (for alternative estimates of income) and even so is careful to declare that uncertainties about short-run movements in velocity imply that only *long-run* growth in income (again, growth rates) should be inferred.

The technique of estimating colonial real income proposed in Chapter 12 (Section 4), and applied to New England, has the virtue of separating long-run and short-run movements in income. However, the technique has its own limitations, some of which are recognized there.

Two other weaknesses of the income variable (Y) generated in Chapter 12 should be addressed. First, of necessity, the variable is synthetic in nature rather than developed from income accounts (which do not exist). However, 'synthetic' need not mean 'unreliable.'

The second issue, noted by Gallman (1999: 23) and Grubb (2004: 352–53), is that non-marketed output (in the form of household production), which is not incorporated in the income variable, was relatively much higher in colonial times than it is today. True, Y pertains only to marketed output; but Y is an index number. Thus it misses the mark on total (market plus non-marketed) output only to the extent that the ratio of marketed to non-marketed output changes over time. There appears to be no reason why this ratio would change systematically in New England during the first half of the 18th century.

Velocity

Velocity is computed in Chapter 12 in the usual way, as a residual to make the equation of exchange hold. An alternative measure of velocity for New England is the inverse of McCallum's (1992: 150) series of 'real money holding per capita.' However, the two velocity series are not comparable, because McCallum (1) operates in a purely long-run context, in effect proxying real income by population, (2) uses the uncorrected rather than PPP-adjusted exchange rate to represent the price level, (3) and employs an inferior money stock series (the unadjusted Brock series).

References

Acworth, A. W. (1925) *Financial Reconstruction in England 1815–1822*. London: King.

Angell, J. W. (1926) *The Theory of International Prices*. Cambridge, MA: Harvard University Press.

Ashton, T. S. (1959) *Economic Fluctuations in England 1700–1800*. Oxford: Clarendon Press.

[*Bank Charter Report*] *Report from the Committee of Secrecy on the Bank of England Charter* (1832), House of Commons, Reports from Committees, vol. 6.

Baxter, W. T. (1945) *The House of Hancock*. Cambridge, MA: Harvard University Press.

Bordo, M. D. (1993) 'The Gold Standard, Bretton Woods and Other Monetary Regimes: A Historical Appraisal', Federal Reserve Bank of St. Louis *Review*, **75**: 123–91.

—— and Schwartz A. J. (1999) 'Monetary Policy Regimes and Economic Performance: The Historical Record', in J. B. Taylor and M. Woodford (eds.) *Handbook of Macroeconomics*, vol. 1A, Amsterdam: Elsevier.

Brock, L. V. (1975) *The Currency of the American Colonies, 1700–1764*. New York: Arno.

—— (1992) 'The Colonial Currency, Prices, and Exchange Rates', *Essays in History*, **34**, University of Virginia.

[*Bullion Report*] *Report on the High Price of Gold Bullion* (1810), House of Commons, Reports from Committees, vol. 3.

Canaan, E. (1925) *The Paper Pound of 1797–1821*. London: King.

Clapham, J. H. (1945) *The Bank of England*, vol. 2. Cambridge: Cambridge University Press.

Clare, G. (1893) *The ABC of the Foreign Exchanges*. London: Macmillan.

Cole, A. H. (1938) *Wholesale Commodity Prices in the United States, 1700–1861*. Cambridge, MA: Harvard University Press.

Collins, M. (1986) 'Sterling Exchange Rates, 1847–80', *Journal of European Economic History* **15**: 511–33.

Coppieters, E. (1955) *English Bank Note Circulation 1694–1954*. The Hague: Martinus Nijhoff.

Davis, A. M. (1900) *Currency and Banking in the Province of the Massachusetts Bay*, Part 1. New York: Macmillan.

Deane, P. (1979) *The First Industrial Revolution*. Cambridge: Cambridge University Press.

Dickinson, W. (1818) *Universal Commerce, or the Commerce of all the Mercantile Cites and Towns of the World*. London: Boosey.

Egnal, M. (1998) *New World Economies*. New York: Oxford University Press.

Einzig, P. (1970) *The History of Foreign Exchange*, 2nd ed. London: Macmillan.

Ernst, J. A. (1973) *Money and Politics in America, 1755–1775*. Chapel Hill: University of North Carolina Press.

Feavearyear, A. (1963) *The Pound Sterling*, 2nd ed. Oxford: Clarendon Press.

Fetter, F. W. (1950) 'Legal Tender during the English and Irish Bank Restrictions', *Journal of Political Economy* **58**, 241–53.

Flynn, D. T. (2001) 'Credit and the Economy of Colonial New England', unpublished thesis, Indiana University.

Gallman, R. E. (1999) 'Can We Build National Accounts for the Colonial Period of American History?' *William and Mary Quarterly*, **56**: 23–30.

Gayer, A. D., W. W. Rostow, and A. J. Schwartz (1953) *The Growth and Fluctuation of the British Economy 1790–1850*, vol. 1. Oxford: Clarendon Press.

Grubb, F. (2004) 'The Circulating Medium of Exchange in Colonial Pennsylvania, 1729–1775: New Estimates of Monetary Composition, Performance, and Economic Growth', *Explorations in Economic History* **41**: 329–60.

Hanson, J. R., II. (1979) 'Money in the Colonial American Economy: An Extension', *Economic Inquiry* **17**: 281–86.

Hawtrey, R. G. (1918) 'The Bank Restriction of 1797', *Economic Journal* **28**, 52–65.

—— (1950) *Currency and Credit*, 4th ed. London: Longmans, Green.

Kelly, P. (1811, 1821) *The Universal Cambist*, vol. 1, London: Lackington.

King, W. T. C. (1936) *History of the London Discount Market*. London: Routledge.

Letwin, W. (1981) 'Monetary Practice and Theory of the North American Colonies during the 17th and 18th Centuries', in B. Bagnoli (ed.) *La Moneta Nell'Economia Europea: Secoli XIII–XVIII*, Florence: Le Monnier.

Marcuzzo, M. C., and A. Rosselli (1990) *Ricardo and the Gold Standard*, trans. J. Hall. New York: St. Martin's.

McCallum, B. T. (1992) 'Money and Prices in Colonial America: A New Test of Competing Theories', *Journal of Political Economy*, **100**: 143–61.

McCulloch, J. R. (1851) *Essays on Interest, Exchange, Coins, Paper Money, and Banks*. Philadelphia: Hart.

McCusker, J. J. (1976) 'Colonial Paper Money', in E. P. Newman (ed.) *Studies on Money in Early America*. New York: American Numismatic Society.

—— (1978) *Money and Exchange in Europe and America, 1600–1775*. Chapel Hill: University of North Carolina Press.

Michener, R. (1987) 'Fixed Exchange Rates and the Quantity Theory in Colonial America', *Carnegie-Rochester Conference Series on Public Policy* **27**: 233–308.

Morgan, E. V. (1943) *The Theory and Practice of Central Banking 1797–1913*. Cambridge: Cambridge University Press.

Murphy, B. (1973) *A History of the British Economy 1086–1970*. London: Longman.
Officer, L. H. (1996) *Between the Dollar-Sterling Gold Points*. Cambridge: Cambridge University Press.
[*Resumption Report*] *Reports Respecting the Bank of England Resuming Cash Payments* (1819), Lord's Committee. House of Commons, Reports from Committees, vol. 3.
[*Report of 1797*] *Report of the Lord's Committee of Secrecy* (1810), House of Commons, Reports from Committees, vol. 3.
Ricardo, D. (1811) *Reply to Mr. Bosanquet's 'Practical Observations on the Report of the Bullion Committee*', in Sraffa (1951).
Shaw, W. A. (1896) *The History of Currency*, 2nd ed. New York: Putnam's.
Silberling, N. J. (1919) 'British Financial Experience 1790–1830', *Review of Economic Statistics* **1**: 282–323.
Smith, B. D. (1985) 'American Colonial Monetary Regimes: the Failure of the Quantity Theory and Some Evidence in Favour of an Alternate View', *Canadian Journal of Economics*, **18**: 531–65.
—— (1987) *Money and Inflation in the American Colonies: Further Evidence on the Failure of the Quantity Theory*, working paper, London, Canada: Department of Economics, University of Western Ontario.
Soetbeer, A. (1879) *Edelmetall-Produktion*. Gotha: Justus Perthes.
Sraffa, P. (ed.) (1951) *The Works and Correspondence of David Ricardo*, vol. 3. Cambridge: Cambridge University Press.
Sumner, S. (1993) 'Colonial Currency and the Quantity Theory of Money: A Critique of Smith's Interpretation', *Journal of Economic History* **53**: 139–45.
Vilar, P. (1991) *A History of Gold and Money*. London: Verso.
Viner, J. (1937) *Studies in the Theory of International Trade*. New York: Harper.
Waterston, W. (1847) *A Cyclopedia of Commerce*. London: Bohn.
Weeden W. B. (1890) *Economic and Social History of New England, 1620–1789*, vol. 2. Boston: Houghton, Mifflin.
Weiss, R. W. (1974) 'The Colonial Monetary Standard of Massachusetts', *Economic History Review*, **27**: 577–92.
West, R. C. (1978) 'Money in the Colonial American Economy', *Economic Inquiry*, **16**: 1–15.

Part IV

Gold

14 Review of *The Monetary History of Gold: A Documentary History, 1660–1999*[*]

A new collection of historical documents is always welcomed by scholars. This is especially so if the collection is theme-oriented, so that specialists can acquire new information or readier access to existing information. Therefore the present volume, presented as a documentary history of gold, warrants careful review. The editor, Mark Duckenfield, explains that the project was financed by the World Gold Council's Public Policy Centre 'to raise awareness among journalists, scholars and the informed public of gold's role as a monetary asset' (p. xvii). This review will be concerned solely with the book's place in the scholarly literature.

What is the World Gold Council? According to its website (http://www.jewelrynet.com/WorldGold/), 'the World Gold Council is a non-profit association of the world's leading gold producers, established to promote the use of gold . . . and its promotional activities cover markets representing some three quarters of the world's annual consumption of gold.' It is interesting, therefore, that the volume contains no documents that deal with gold production, gold consumption, or gold investment. Indeed, these categories do not even enter the index, which, at over eleven pages, is not short. Only in one editorial passage does the editor discuss the role of gold as an investment. Duckenfield sees gold and the stock market as investment alternatives, and observes that 'the stock market's decline since 1999, as well as the increase in international tensions as a result of the terrorist attack on the World Trade Center in New York on 11 September 2001 and the second Gulf War in 2003, brought a resurgence in the gold market as investors returned to gold as a traditional safe-haven in troubled times' (p. 327). He explicitly states that 'gold . . . [is] a non-interest bearing asset.' Notwithstanding the sponsorship, this is a book that adopts a scholarly viewpoint and can be appreciated by historians.

In this review the book is placed in context with existing collections of historical documents that have the same, gold-oriented, theme – in whole or in part. These books are listed in Table 14.1, in chronological order; full citations are provided in the references at the end of this review.

In terms of absolute size Duckenfield is in the middle of the pack, fourth out of seven in total pagination. With 100-percent devotion to documents,

Table 14.1 Pages devoted exclusively to historical documents

Book	*Pages*	
	Number	*Percent of book*
U.S. Senate (1879)	596	65
Horton (1887)	87	28
Laughlin (1896)	38	11
Huntington and Mawhinney (1910)	812[a]	100
Shrigley (1935)	87	81
Krooss (1969)	3232[a]	100
Duckenfield (2004)	536[a]	100

a Includes index.

the book's relative size in this respect is shared by only two other collections: Huntington and Mawhinney (subsequently referenced as 'Huntington') and the massive, four-volume work of Krooss. Considering the other books in Table 14.1, what else is in them apart from historical documents and associated editorial commentary? U.S. Senate contains proceedings of the International Monetary Conference of 1878. These are presented as contemporary recording rather than historical documentation; therefore these proceedings are not included in the historical-document pagination. Horton and Laughlin are each mainly a monetary treatise. Shrigley offers non-document information, quantitative and qualitative, on gold and the Bank of England (of which, more below). It may be noted that S. Dana Horton, a U.S. delegate to international monetary conferences, not only authored and edited the historical documents in Horton (1887) but also had been responsible for the historical documents included in U.S. Senate (1879).

Table 14.2 summarizes the anatomy of Duckenfield and the six other works. In Table 14.1 all historical documents are included to obtain the page count; in Table 14.2 only those documents that fall within categories covered by Duckenfield are incorporated. (These are called 'pertinent documents,' in this review.) Some judgment on the part of this reviewer is involved. For example, Duckenfield has one document pertaining to the greenback period as a suspension of specie payments; documents under that category are included for the other works. An opposite example: Duckenfield has a document on goldsmith banking; this reviewer judges that insufficient to incorporate the category 'commercial banking.' Again, measured by number of documents, Duckenfield is fourth among the seven collections.

All the collections except Shrigley are divided into sections, some chronological, some by country, some by topic, as Table 14.2 shows. Only two of the works have systematic subsections. Within sections and subsections, ordering is uniformly chronological. Regarding editorial commentary, Duckenfield is unique, and deserves praise, for having editorial introductions in all three manifestations: for the entire volume, by section, and for the individual

Table 14.2 Anatomy of Duckenfield and related works

Book	*No. of docs.*[a]	*Sections*			*Editorial commentary: by*		
		No.	*Basis*	*Subsections*	*Vol.*	*Section*	*Document*
U.S. Senate	91	5	country[b]	no[c]	yes	no[d]	no
Horton	40	2	topic	no	no	no	no[e]
Laughlin	22	6	country[f]	no	no	no	no
Huntington	142[g,h]	4[i]	topic	yes[j]	no	no	no
Shrigley	20	1			yes	no	no
Krooss	123[h]	10	time	yes[k]	yes[l]	yes	no
Duckenfield	153	3	time	no	yes	yes	yes

a Of type in Duckenfield volume.
b Also 'Miscellaneous' and 'Monetary Union.'
c Except for one subsection.
d Except for one section.
e Except for one document.
f Also 'Latin Monetary Union.'
g 'Acts' only. Excluded are 'Revised Statutes,' of which 162 are pertinent.
h Excluded are documents in categories omitted by Duckenfield: government financing, commercial banking, U.S. central banking, U.S. paper money (except for greenback period).
i Of which only two are of type in Duckenfield volume.
j 'Acts' and 'Revised Statutes.'
k Basis: by topic.
l By Paul A. Samuelson.

documents within each section. Because entries in a collected volume of documents generally are excerpts rather than the entire documents, this reviewer appreciates Duckenfield's practice of calling attention to non-reprinted parts of documents.

The three sections of Duckenfield warrant discussion, here in the context of editorial commentary. The introduction to the first, 'The Rise of the Gold Standard, 1660–1819,' is concerned entirely with British monetary history. Duckenfield observes England's movement from bimetallism to a de facto gold standard in 1717. He notes the interruption of the Bank Restriction Period in this process. It is reasonable to confine discussion to Britain, because it was the only country on a gold standard well into the nineteenth century. However, it would have been in order to discuss the bimetallist systems of other countries.

The second section, 'The Heyday of the Gold Standard, 1820–1930,' has a broader introduction, including topics such as the expansion of the gold standard, the price specie-flow mechanism, the U.S. shift from an effective silver to an effective gold standard (with the interruption of the greenback period), the deflation of 1873–1896, and London as the center of the gold standard. The end of the classical gold standard with World War I is noted, as is the return to the standard after the war. Duckenfield discusses the issue of convertibility but can be criticized for ignoring that of credibility (of countries' commitment to convertibility at the existing mint price), which underlay the success of the classical gold standard.

In his introduction to the third section, 'After the Gold Standard, 1931–1999,' Duckenfield sees a weak institutional structure as the cause of instability of the interwar gold standard. 'Domestic social tensions' and the 'prospect of substantial budget deficits' drove countries off the gold standard. War and 'new social realities' meant that political and economic institutions that supported the gold standard could not overcome political demands that occurred during the Great Depression. Again, reference to the issue of credibility, now the lack thereof, in government's commitment to convertibility would have been in order.

The introduction also discusses the International Monetary Fund, the role of the dollar, and the 'Triffin dilemma' (the trade-off between liquidity and confidence). On the U.S. suspension of gold convertibility in 1971, Duckenfield writes: 'Ironically, although it was the weakness of the dollar relative to gold that brought about the collapse of the Bretton Woods system, it was gold that was removed from its primary position as a monetary asset while the devalued dollar became even more crucial to the smooth operation of the international economy' (pp. 326–327). It can be argued, rather, that it was U.S. commitment to a fixed dollar price of gold that artificially made gold the first-class monetary asset.

Table 14.3 divides the pertinent documents of each work into chronological

Table 14.3 Number of pertinent documents – by type and time period

		No. of documents	
Book	*Period*	*Official*	*Private*
U.S. Senate	1666–1819	46	5
	1820–1877	36	4
Horton	1663–1817	26	5
	1818–1882	9	0
Laughlin	1778–1819	2	0
	1820–1893	20	0
Huntington	1778–1819	20	0
	1820–1908	122	0
Shrigley	1694–1819	2	0
	1820–1930	14	2
	1931–1933	1	1
Krooss	1652–1819	15	0
	1820–1930	64	7
	1931–1968	35	2
Duckenfield	1660–1819	34	18
	1820–1930	34	2
	1931–2004[a]	61	4

a Although the title of the book states the full period as 1660–1999, one document (in an appendix) pertains to the year 2004.

sections (pre-1820, 1820–1930, post-1930) corresponding to the Duckenfield partitioning, except that Horton's original partitioning is retained, because it is so close to that of Duckenfield chronologically. Of course, the four documents antedating Shrigley lack the third (post-1930) section, because of the date of publication.

What are the components of official versus private documents? Official documents consist of country and international items. Country official documents include acts, resolutions, announcements, reports, memoranda, communications, statements, declarations, representations, speeches, notes on petitions, parliamentary diaries, proclamations, mint correspondence, and press conferences. International documents (all official) consist of treaties, conventions, resolutions, agreements, press releases, communiqués, and decisions. Private documents include treatises, books, pamphlets, diaries, discourses, petitions, speeches, reports, correspondence, memoirs, newspaper articles, and memoranda.

As one would expect in collections of documents, the vast majority of documents are official, in all the works. Duckenfield is unique in having private documents constitute a significant proportion – over one-third the total number – of documents in the pre-1820 period.

Table 14.4 offers an alternative division of pertinent documents – by country (Britain, United States, other countries), with international as a separate category. For private documents, the subject country is taken. For official documents, the country category is the country of the official document

Table 14.4 Number of pertinent documents – by country and time period

		Number of documents			
Book	*Period*	*Britain*	*United States*	*Other countries*	*International*
U.S. Senate	1666–1819	26	20	5	0
	1820–1877	7	17	2	14
Horton	1663–1817	31	0	0	0
	1818–1882	5	0	0	4
Laughlin	1778–1819	0	1	1	0
	1820–1893	0	13	4	3
Huntington	1778–1819	0	20	0	0
	1820–1908	0	122	0	0
Shrigley	1694–1819	2	0	0	0
	1820–1930	16	0	0	0
	1931–1933	2	0	0	0
Krooss	1652–1819	0	15	0	0
	1820–1930	0	71	0	0
	1931–1968	0	37	0	0
Duckenfield	1660–1819	50	1	1	0
	1820–1930	22	7	5	2
	1931–2004	27	23	4	11

rather than the subject country or countries. All seven works concentrate on Britain and/or the United States. Huntington and Krooss deal only with the United States, Shrigley only with Britain. Laughlin has U.S., other-country, and international – but not British – documents; while Horton includes only British and international documents. The only works with documents in each category are the earliest and latest: U.S. Senate and Duckenfield. As would be expected, U.S. Senate has somewhat more U.S. than British documents overall (but not for the pre-1820 period).

One would predict a balanced British/U.S. division on the part of Duckenfield, given that neither the book-title nor the sponsor is specific-country oriented. Then one would be disappointed, because a British emphasis is present in every period. The asymmetry is apparent in several ways:

1. There are nine documents of the Bank Restriction Period but only three from the greenback period and nothing on other U.S. suspensions of specie payments.
2. There is an entry for the Bank of England charter, but not for the Federal Reserve Act.
3. There are more entries for Acts of Parliament than for U.S. legislation.
4. The British Coinage Act of 1870 is reprinted in full; not so the U.S. Coinage Act of 1873, which admittedly is a longer Act.

In fairness to Duckenfield, it should be noted that, while Winston Churchill's famous Budget Speech of 1925 returning the United Kingdom to the gold standard is excerpted, William Jennings Bryan's at least equally famous 'Cross of Gold' Speech is reprinted in full. Also, there are many entries involving U.S. abandonment of the gold standard in 1933–1934.

Notwithstanding the generally British orientation of Duckenfield, Horton is the better source for material on the history of the guinea – perhaps the most famous coin in British history, and, along with the (new) sovereign introduced in 1817, one of the country's two most important coins. Only seven of Horton's 31 documents on the guinea are included in Duckenfield. The guinea is notable as a coin for two reasons. First, its initial value of 20 shillings corresponded to the pound sterling. Interestingly, the guinea was not the first coin with this property; that distinction belongs to the old sovereign, introduced in 1489. Second, the fineness of 11/12th was firmly established with the guinea (and continuing with the new sovereign); but again the guinea was not the first coin with that fineness (that honor belonging to the crown in 1526).

Shrigley, totally specialized on Britain, has a specific theme within gold. She writes: 'The purpose of this collection of documents is to show the official position of gold as a marketable commodity from the Incorporation of the Bank of England to the Gold Standard (Amendment) Act of 1931' (p. vii). Of her 20 documents, 12 are not in Duckenfield.

Table 14.5 breaks down the 'other-countries' category of Table 14.4 into

Table 14.5 Number of pertinent documents – by 'other country' and time period

Book	*Period*	*Number of documents* France	Germany	Switzerland	Remaining
U.S. Senate	1666–1819	5	0	0	0
	1820–1877	0	2	0	0
Laughlin	1778–1819	1	0	0	0
	1820–1893	0	2	0	2[a]
Duckenfield	1660–1819	1	0	0	0
	1820–1930	1	1	2	1[b]
	1931–2004	1	0	2	1[c]

a Italy, Austria.
b Chile.
c Yugoslavia.

specific countries. Duckenfield does not provide a rationale for his concentration on Switzerland in the 1920–1930 and post-1930 periods in this respect, nor for inclusion of material on countries such as Chile and Yugoslavia post-1930. It is also arguable that France and Germany deserve greater attention than all three works give these countries.

Table 14.6, similarly, partitions the 'international' category of Table 14.4. Duckenfield can perhaps be criticized for neglecting the international monetary conferences of the nineteenth century. Yet he deserves praise for including the, post-World War I, Treaty of Versailles – relevant because of the gold-denomination of the monetary obligations imposed on Germany.

Duckenfield deserves praise on a number of counts. First, for some documents, the contents of appendices are listed. (Indeed, that is sometimes the full text of the entry.) These contents can be useful references for the scholar. Second, Duckenfield makes use of generally neglected sources: Bank of England archives and the House of Lords Record Office. Third, some documents may be new to historians. Examples: a 'confidential telegram' (one of many) sent on September 20, 1931, from the Bank of England to domestic and foreign correspondents; the Rothschild letter on fixing the price of gold in 1939, just prior to World War II.

A serious limitation of the Duckenfield volume is the neglect of quantitative information. Only three documents have a quantitative aspect: the original Articles of Agreement of the IMF, which contains the list of country quotas; the IMF Executive Board decision on the Smithsonian Agreement, which lists exchange rates for member countries; and, perhaps most interesting to historians because probably not available elsewhere, three documents from the Bank of England's Archives providing data on gold holdings of countries occupied by Germany, in 1940. In contrast, the documents in U.S. Senate contain many useful tables on U.S. exchange rates, gold and silver prices, and coinage.

Table 14.6 Number of pertinent international documents – by organization and time period

Book	*Period*	*Number of documents*					
		International monetary conferences	*Latin Monetary Union*	*Bank of International Settlements*	*International Monetary Fund*	*Central banks*	*Other*
U.S. Senate	1820–1877	12	2				
Horton	1818–1882	4					
Laughlin	1820–1893		3				
Duckenfield	1820–1930		1				1[a]
	1931–2004			2	4	3	2[b]

a Treaty of Versailles.
b Tripartite Agreement, Smithsonian Agreement.

Shrigley presents several useful time series (which are not included in the list of her documents, in Tables 14.2–14.4): the annual gold–silver market price ratio, 1867–1932; the London market price of gold, annual 1870–1932, daily 1919–1925; and the London market price of silver, monthly 1833–1933. The last is an insert at the end of the book, and includes also annual data on silver coined in England, the amount of bills and telegraphic transfers drawn in England on Indian governments, exports of silver to the East, imports of silver, average Bank Rate, and remarks (generally historical). It is a large and impressive table, which, unfortunately, because not attached to the volume, may be missing from many copies. Not a time series, but nevertheless useful, is a list of Governors of the Bank of England from inception to 1920, along with dates of service.

In conclusion, the Duckenfield volume is a useful addition to collections of historical documents on gold, and would be best utilized by scholars in conjunction with existing works of a similar ilk.

Note

* Originally published in Lawrence H. Officer, 'Review of Mark Duckenfield *The Monetary History of Gold: A Documentary History, 1660–1999*,' Economic History Services, Dec 14, 2004, URL : http://eh.net/bookreviews/library/0878.shtml.

References

Horton, S. Dana (1887). *The Silver Pound and England's Monetary Policy since the Restoration, together with the History of the Guinea, illustrated by contemporary documents*. London: Macmillan.

Huntington, A. T., and Robert J. Mawhinney, eds. (1910). *Laws of the United States Concerning Money, Banking, and Loans, 1778–1909*. National Monetary Commission, Senate Document No. 580, 61st Congress, 2nd session. Washington: Government Printing Office.

Krooss, Herman E., ed. (1969). *Documentary History of Banking and Currency in the United States*, four volumes. New York: Chelsea House in association with McGraw-Hill.

Lauglin, J. Laurence (1896). *The History of Bimetallism in the United States*. New York: D. Appleton.

Shrigley, Irene, ed. (1935). *The Price of Gold: Documents Illustrating the Statutory Control through the Bank of England of the Market Price of Gold, 1694–1931*. London: P.S. King & Son.

U.S. Senate (1879). *International Monetary Conference . . . held . . . in Paris, in August 1878, under the auspices of the Ministry of Foreign Affairs of the Republic of France*. Senate Executive Document No. 58, 45th Congress, 3rd session. Washington: Government Printing Office.

15 Reserve-asset preferences in the crisis zone, 1958–67*

This study provides an econometric analysis of the reserve-asset behavior of the outer countries over a critical time span of the international monetary system, between 1958, the year the system entered its 'crisis zone' (that stage in which the probability of collapse of the system became apparent to the participants), and 1967, when agreement was reached on the SDR reform of the system. It tests for changing reactions of these countries with respect to the confidence problem as the system moved further into the crisis zone. Such an altered behavior pattern will be denoted as 'restraint' if it involves the holding of more dollars and fewer other reserve assets (primarily gold) than a country otherwise would do. This restraint has been postulated in theoretical studies by Officer and Willett [14, 15], but thus far their work has not been subjected to empirical testing.

How can one determine the beginning of the crisis zone of the international monetary system? Either a stock or flow criterion may be used. The stock definition of the crisis zone was first presented in a theoretical study by Kenen [11]: it is the period in which the U.S. reserve ratio (official reserve assets to liquid liabilities to all foreigners) is below unity. This ratio exhibits a declining trend from 1951 to 1967, a trend which falls below unity in 1960. A second definition of the crisis zone is implicit in the related work of Mundell [13][1] and was adopted in the later study of Officer and Willett [15]: the crisis zone is the period in which the outer countries view the United States as having an undesirably large balance-of-payments deficit. The first year of continuing substantial U.S. deficits would be 1958, and this is taken as the start of the crisis zone.

I. An analysis of the confidence problem

The keynote of this study is *power*. On the one hand, the power of a country represents its ability to pursue its own reserves-management policy, more or less independent of pressures (self-imposed or external) placed on it to preserve the system. On the other hand, an ideal indicator of any restraint in this policy would be a decreased influence of the country's power on the management of its international-reserves portfolio. The power thus far discussed refers to that of the outer countries, meaning all countries except the center country, the United States. The United States, on its part, uses power to influence the outer countries to exhibit restraint, to limit gold conversion of their dollar holdings. It has the ability to use both 'threats' and 'bribes' for this purpose. The concepts of power, threat, and bribe present a measurement problem; and definition of the variables to represent them is an important feature of the study.

The U.S. unilateral suspension of the gold-convertibility status of the dollar in August 1971 represented an ultimate kind of threatening action. With the gold window closed, the confidence problem expressed itself solely in the form of negotiations to reform the system – negotiations that resulted in agreement four months later. The present study is concerned with the outer countries' reserve-asset behavior during an earlier period, one in which the gold-conversion right of foreign officially owned dollars had not been abrogated but during which a reform of the system was negotiated. The period begins when the system entered the crisis zone (1958), includes negotiations for reform which began in November 1963, and ends with adoption of the Special Drawing Rights reform (1967).

The theory underlying this study is based on the Officer–Willett model of the international monetary system. The system is viewed as a game with two players, or, more precisely, a unitary player (the United States) on one side and an amalgam of players (all other members of the International Monetary Fund, the outer countries) with more or less decentralized decision-making on the other. All countries have a mixed-motive behavior pattern – to gain national advantage and to preserve the system. In the case of the outer countries, these two motives may become opposing forces in connection with their decisions regarding the composition of their international reserves, in particular, how many dollars to hold in relationship to gold. National advantage may dictate a lower dollar/gold ratio and preserving the system a higher one. At the same time the United States may take steps to induce countries to maintain this ratio higher than otherwise.

The Officer-Willett model of the international monetary system (like that of Mundell) involves the outer countries' use of gold conversions to signal the United States to correct its balance-of-payments deficit. However, an inherent part of the model is the fact that each country has a stake in the system, i.e., would suffer a disutility if the system collapsed. 'The loss of gold in significant amounts would tend to reduce the *u* [stake in the system] of the center country. . . . One event that could occur under these circumstances would be a U.S. decision to opt out of the system' [15, p. 54]. It is one of the splendid paradoxes of the present system that while gold is undervalued with respect to the dollar, gold would probably be found to have been *overvalued* if the system collapsed through demonetization of gold by the United States. 'The reserve center's response to a run on its currency might well be repudiation of gold, which could lead to its general demonetization and a *fall* in its price. A substantial shift in expectations concerning the U.S. response to a run on the dollar thus could lead to a marked change in the outer countries' subjective evaluations of the desirability of dollars and gold as reserve assets'[2] [14, p. 691]. Thus, even from the narrow standpoint of the riskiness attached to dollar assets, there is reason for restraint in gold conversion as the U.S. reserve ratio deteriorates.

A country's stake in the system, moreover, has greater underpinnings than even an enlightened portfolio-management policy. It 'reflects the attitudes of official decision-makers as to "the national interest" of their countries in the preservation of the system' [15, p. 48]. Trading and other international economic relationships, international political and military alliances, and domestic economic and political considerations would all be involved in determining a country's stake in the system.

Its stake in the system gives the outer country an incentive to show restraint in gold-conversion policy, in other words (and in the context of the present study), to hold a higher dollar/gold ratio than it would otherwise. This behavior will be especially prevalent as the probability of the system's collapse increases. And the U.S. reserve ratio is an indicator of this probability. Thus, paradoxically, the system may become stronger as it appears weaker! The situation is even more complicated: even though all wish to preserve the system, there is obvious incentive to throw the burden of preservation onto another 'player.' Thus there is scope for behavior of a bluffing, bribing, or threatening nature in efforts to accomplish this end.

Essentially political factors determine an outer country's reserve-asset composition in the crisis zone. On the one hand, countries would like to reduce their ratio of dollar holdings to total reserve assets (dollars/reserves) – because of (1) reduced confidence in the U.S. ability or willingness to support the dollar in the event of a run on the dollar and (2) a desire to signal the U.S. to correct its deficit. On the other hand, countries will keep a higher dollars/reserves ratio than otherwise, (1) just to preserve the system and (2) as a reaction to bribes or threats directed against them by the United States. These bribes and threats need not be explicitly made. They are apparent in the power that the United States can exercise with respect to the country in question. Of course, that outer country also has some power. Its own power represents its ability to withstand threats and ignore bribes. However, because of increasing concern for the endurance of the system, the use of this power is expected to decline over time, as the system moves further into the crisis zone.

The hypotheses to be tested, therefore, are as follows.

1. At any given time in the crisis zone, an outer country's dollars/reserves ratio will be higher (*a*) the greater its susceptibility to the U.S. use of power in the form of bribes or threats and (*b*) the less its own power, which enables it to counteract such U.S. pressure.
2. As the system proceeds deeper into the crisis zone, (*a*) there is an increased apparent susceptibility of the outer country to U.S. bribes and threats as the United States applies pressure more intensively and

(*b*) there is a decrease in the outer country's resort to its own power to withstand U.S. pressure for a high dollars/reserves ratio.

The research strategy employed is to relate the foreign-exchange/reserves ratio – ideally this would be the dollars/reserves ratio – to measures of a country's power and of its susceptibility to bribes and threats exercised by the United States. Cross-section multiple regressions are estimated in which foreign-exchange/reserves is the dependent variable. This analysis is performed at ten points in time (1958–1967, with end-of-year data on the dependent variable). For each point in time, the sample is the largest number of countries for which the required data exist. This moving cross-section approach provides the opportunity of examining and comparing regression coefficients not only at a given point in time but also over time. It is the latter property which enables one to test for restraint in the dollar-gold behavior of the outer countries.

II. The variables

Dependent variable

The dependent variable, ideally, should be the ratio of a country's official dollar holdings to its total international reserves. Unfortunately, data on the numerator are not available, except in rare cases.[3] If one accepts the I.M.F. definition of liquidity, then the denominator exists in published form with the following breakdown: reserves ≡ gold + foreign exchange + position in the Fund. Symbolically,

$$R \equiv G + E + F \tag{1}$$

Now, foreign exchange would have three components: foreign exchange ≡ dollars + pounds + other currencies.

$$E \equiv D + P + O \tag{2}$$

While we do not have the breakdown in (2), it is reasonable to assume that, typically, other currencies O are negligible in countries' reserves portfolios.[4] Of course, such a statement cannot be made about British pounds P, which are the preponderant part of the foreign-exchange reserves of countries within the sterling area and may also be held, in lesser proportions, by countries outside the area.

In addition to foreign exchange E, there exists another data series an unknown component of which is dollar holdings D. The United States reports dollar assets held by individual countries, but does not separate official from private holdings. For many countries, however, the United States provides a series in which nonbank private holdings are netted out from the total. Thirty-five of the 76 countries in the sample have this property; for such countries, privately held dollars are exclusive of nonbank holdings in the identity that follows: total dollars ≡ officially held dollars + privately held dollars.

$$D_t \equiv D + D_p \tag{3}$$

Now, considering the variables in identities (1), (2), and (3), we know R, G, E, F, P, D_t, and are prepared to take O as zero. The variable we should like to have is the ratio of officially held dollars to total reserves (D/R), but officially held dollars D are not known. However, identities (2) and (3) show that D is a component of two known variables, namely, officially held foreign exchange E and officially held *plus* privately held dollars D_t. Thus we have two proxy measures of D/R: the foreign-exchange/reserves ratio (E/R) and the total-dollars/reserves ratio (D_t/R)[5]. Which is the better proxy?

The answer, of course, depends on the correlations of official foreign exchange E to official dollars D and of total dollars D_t, to official dollars D, where all variables are deflated by total reserves R. According as corr. (E/R, D/R) is greater (less) than corr. (D_t/R, D/R), then E/R (D_t/R) is the appropriate proxy of the two. The fact that all of E, while only a part of D_t, is a component of R establishes no presumption in favor of corr. (E/R, D/R), just as the fact that D_t consists only of dollar (and not pound or other-currency) holdings implies none in favor of corr. (D_t/R, D/R). With D unknown, the correlation coefficients cannot be calculated – if D were available, a proxy for it would be unnecessary – and no definite answer as to the better measure can be established.

Nevertheless, there are two tests that can be used, with substantial justification, to choose one measure over the other. The first consists of a comparison of the distributions of the variables E/R and D_t/R. Mere inspection revealed that D_t/R is an extremely volatile variable compared to E/R. In the 1967 sample, for example, the ratio of the standard deviation to the mean of D_t/R is more than four times that of E/R. Such a test, of course, is not sufficient to choose E/R as the proxy. It is conceivable, though very unlikely, that the greater variability of D_t/R serves better to capture the true variation of D/R.

The obvious second (and complementary) test is to run the regression analysis using E/R and D_t/R as alternative explanatory variables. Experimentation showed that much more meaningful and significant results were obtained with E/R rather than D_t/R. Given that the hypotheses have some basis in reality, it is only logical to interpret the second test as establishing the fact that the substantial volatility of D_t/R is indigenous to itself and not a reflection of the variation in the series D/R. Therefore the foreign-exchange/reserves ratio (E/R) was chosen as the dependent variable for the analysis that follows; E/R = ratio of foreign exchange to total reserves, end of year, percent [9].[6]

Explanatory variables

Four groups of explanatory variables are considered: outer country's power, economic dependence on the United States, U.S. foreign aid, and U.S. diplomatic activity. As there are several alternative and/or complementary variables in each group, they were the subject of experimental regressions, the final results of which are reported in Section III. The discussion here describes which variables survived the experimentation process and which were excluded from the final regression equations. The number in parentheses following the initial listing of each variable is the correlation coefficient of that variable with foreign-exchange/reserves. To be precise, it is the mean of the correlation coefficients calculated for each of the ten samples (1958–67) rather than the overall mean computed from the pooled data. A superscript of a single asterisk indicates significance at the five-percent level; a double asterisk indicates significance at the one-percent level.

Finally, the time dimension of variables should be noted. The dependent variable is an end-of-year figure, while all 'flow' explanatory variables are annual, referring to the year in question. Information on 'stock' explanatory variables is given as such variables are introduced.

Power

To measure the power of a country is by no means a straightforward task. There are many variables which, singly or in combinations, can be used to represent this power. One must resort to experimentation with alternative groups of variables coupled with various alternative and/or complementary variables within each group. Two such groups were explored: one consisting of unidimensional (or 'intensive') measures; the other composed of multidimensional (or 'extensive') measures, those that are combinations of unidimensional variables. Though power is a relative concept, it can be measured absolutely for the purpose of this study. The reason is that the power of an outer country is to be considered in opposition to the power of the center country, the former measuring, as it does, the country's ability to withstand threats and ignore bribes exercised by the United States. The purist would deflate any power measure by the corresponding value for the United States, but, of course, such a procedure would not change the results.

A. Intensive measures

One may say that power is 'political' or 'social,' which merely means that it represents one country's ability to influence, affect, or control another – in the context of this study, the ability to *withstand* such influence, effect, or control. Yet a *single* variable measuring this power must surely be economic or military in nature. Thus the following three such measures were tried as explanatory variables:

GNP (–.36)** = gross national product, millions of dollars [9, 18, 20]
MIL (–.36)** = military expenditure, millions of dollars [19, 20]
BOMBS (–.23) = nuclear potential, number of bombs per annum [3, 5]

In experimental regressions MIL gave the best results of the three variables; so it was selected as the 'intensive' measure of power. GNP was less strong and significant a variable than MIL in regressions, although it shared the same general properties. BOMBS, on the other hand, which was tried as an explanatory variable supplementary to MIL or GNP, failed completely. There are perhaps two reasons for the poor performance of BOMBS. The more important is the large number of zero observations: a positive value exists for only 14 of the 76 countries in the sample. The second reason is errors in the variable, which have two sources. First, the nuclear potential (or actuality, in the case of the United Kingdom and France) can only be estimated, and the margin for error surely is great. Second, except for the actual nuclear powers (the United Kingdom and France), only a single estimate of nuclear potential for any country could be found, rather than one which changed over time. Were it not for errors in measuring the variable, one would think that BOMBS might have functioned as a good supplement to MIL in the regressions; for BOMBS measures potential, as distinct from actual, military capability. Indeed, it provides an indication of the military research-and-development capacity of the country in question.

B. Extensive measures

If one were forced to choose a unidimensional measure of power, presumably it would be an economic or military variable, as selected above. Yet power, or rather the quantitative measurement of it, is certainly a nebulous concept. This suggests that a superior measure would be multidimensional in nature, one based on a large number of measures of a country's economic, military, political, and social status. A multivariate analysis would be required to combine these underlying variables into the desired measures of power.

Fortunately for our purposes, this task has already been performed. Bruce M. Russett [16] extracted five factors from 54 economic, military, political, and social variables in 82 countries. He found reason to label these factors 'economic development' (accounting for 31 percent of the total variation), 'communism' (11 percent), 'intensive agriculture' (6 percent), 'size' (7 percent), and 'Catholic culture' (5 percent). Russett transformed the underlying variables to a uniform zero-to-one scale[7] and took principal components, thus yielding factor scores by country. The factor scores were standardized with zero mean and unit standard deviation. Russett showed results for four factors (excluding 'size'). These factors may be interpreted as broad or extensive measures of power compared to the intensive ones discussed above, and they were tried as explanatory variables.[8] Their correlation with the dependent variable are as follows: 'economic development' (–.22), 'communism' (.08), 'intensive agriculture' (–.13), and 'Catholic culture' (–.27).*

Only two of the variables were significant in the regressions, namely, 'economic development' and 'Catholic culture,' the factors which have the highest correlation with the dependent variable. The sign of the correlation coefficient of 'economic development' is negative as expected, as this variable is a measure of power comparable to GNP. However, what interpretation can one put on the negative effect (and significance) of the factor labelled 'Catholic culture'? To answer the question, consider the factor scores (in parentheses) of the United States on the four principal components: 'economic development' (1.90), 'Catholic culture' (0.54), 'communism' (–1.11), 'intensive agriculture' (–0.87) [16, p. 28]. The positive sign of the indexes of both 'economic development' and 'Catholic culture' suggests that 'Catholic culture' can be interpreted as a complementary measure of (the concept) economic development. Supporting evidence is the high loadings on 'Catholic culture' (second only to 'economic development') of GNP per capita, radios per 1,000 population, television sets per 1,000 population, speakers of the dominant language as percent of the population, and total population (negative coefficient). The reason for the designation 'Catholic culture' is that this factor is first (and, not unexpectedly, 'economic development' second) in Catholics as percent of the population and Christians as percent of the population [16, pp. 17–21]. Thus, 'economic development' and 'Catholic culture' are used as complementary measures of power in the analysis, and therefore are relabelled as EC DEV 1 and EC DEV 2, respectively.

What can one say of the two factors that were insignificant in regressions? The correlation of 'communism' is positive, whereas one would expect 'communist' countries to have low dollars/reserves ratios. The reason for this anomaly is the exclusion of the communist-bloc states from the sample for this study,[9] whereas such countries were included in Russett's analysis. 'Communist' attributes probably are scattered randomly among nearly all the countries in the sample. The low correlation coefficient of 'communism' (lowest among all factors) attests to the nonapplicability of this variable for the present study.

The remaining factor, 'intensive agriculture,' is negatively correlated with foreign-exchange/reserves. It is hard to see why such a variable should have an influence either way, representing, as it does, the state of agriculture and the density of population of a country. Presumably, countries that load high on this factor (e.g., Belgium, Netherlands, South Korea, Taiwan) are unlike the U.S. in certain respects, but one fails to see why dollar-gold behavior should be affected other than randomly. Indeed, the low correlation coefficient of 'intensive agriculture' and its insignificance in the regressions bear out this interpretation.

Economic dependence on U.S.

Two alternative measures of an outer country's economic dependence on the United States were tried as follows:

TRADE/TRADE$_t$ (.31)* = ratio of country's trade with the U.S. and U.K. to its total trade, percent [10]
TRADE/GNP (.32)* = ratio of country's trade with the U.S. and U.K. to its gross national product, percent [9, 10, 18, 20]

Trade is defined as the sum of merchandise exports and imports. Because the numerator of the dependent variable is all foreign exchange, the nondollar part of which is predominantly British pounds, the country's dimension of trade is not trade with the United States alone but trade with both the United States and the United Kingdom. This technique allows one to avoid eliminating countries from the sample, whereas in a related study Hagemann [8] dropped the sterling-area countries even though he followed the same technique. (His corresponding variable is the ratio of a country's exports to the United States and the United Kingdom to its total exports.) The irony is that Hagemann lapses into the interpretation of his dependent variable (identical to the one in the present study) as representing *gold* (rather than *dollar*) behavior. Therefore he includes trade with both reserve-currency countries as affecting gold versus foreign-exchange behavior. The present study is concerned with the *dollar* behavior of outer countries – dollars versus gold plus other reserve assets. Therefore trade with the United Kingdom is included in the numerator of the trade variables only to compensate for the error in the dependent variable, that is, the use of foreign exchange as a proxy for dollars.[10]

The two trade measures, TRADE/TRADE$_t$ and TRADE/GNP, were tried as alternative explanatory variables in regressions. Although in terms of the correlation with the dependent variable there is little to choose between them, TRADE/GNP gave the better results in the regressions. TRADE/GNP is a more 'inclusive' economic variable, representing the role of the United States in the country's *production*, whereas TRADE/TRADE$_t$ has a restrictive connotation, limiting its scope to the role of the United States in the country's *trade*. In line with the present analysis, as distinct from that of Hagemann, the trade variable (TRADE/GNP in this study) does *not* reflect solely or primarily balance-of-payments considerations.[11] Rather, it represents *economic dependence* on the United States and hence the likelihood of the country's holding more dollars due to this dependence. The fear is not just harm to trade but harm to production as a result of the U.S. action or inaction. The trade variable measures susceptibility to threatening behavior applied by the reserve center, the United States.

U.S. foreign aid

Foreign aid in general, and particularly that of a bilateral nature, plays the role of a 'bribe' on the part of the donor country toward the recipient country. In the context of the international monetary system, the U.S. aid received by an outer country constitutes a leverage on that country which the United States can use to persuade it to hold a higher dollars/reserves ratio than otherwise. In correspondence with the alternative economic (GNP) and military (MIL) measures of power, two kinds of foreign-aid variables were tried: (1) total aid (economic plus military) deflated by gross national product to correct for country size and (2) military aid deflated by military expenditure. Furthermore, two alternative concepts of military aid were used, the one narrowly defined and the other inclusive of U.S. defense expenditures abroad. Some such expenditures actually are additional military aid, but most of them are simply expenditures on goods and services for the use of the U.S. military. In the latter case the transactions represent a contribution to the economy of the host country which is directly related to U.S. military activities, which suggests that it is not unreasonable to add these expenditures to military aid received. Thus the three alternative explanatory variables are:

AID/GNP (.27)*	= ratio of U.S. economic and military aid to gross national product, percent [1, 9, 18, 20]
AID_m/MIL (.15)	= ratio of U.S. military aid to military expenditure, percent [1, 19, 20]
(AID_m + DEA)/MIL (.14)	= ratio of U.S. military aid plus U.S. defense expenditures in the country to military expenditure, percent [1, 19, 20, 21]

The regression results were in accordance with the correlation coefficients of the three variables; AID/GNP performed best, and it was selected as the variable to measure U.S. aid.

U.S. diplomatic activity

U.S. diplomatic activity in a country might be a good indicator of the intensity of political pressure that the United States is applying to that country; in the context of the present study, this would be pressure to hold a higher dollars/reserves ratio than the country would do otherwise. Then how can the intensity of U.S. diplomatic activity be measured quantitatively?

One measure would be simply the number of U.S. diplomats assigned to the country in question. There appears to be a problem of availability of the relevant information here, although corresponding aggregate data (total numbers of diplomats sent and received by each country) have been analyzed for one point in time by Alger and Brams [2]. A measure that is probably superior because of its greater variation is the civilian employment of the

U.S. government located abroad. Again, two alternative concepts were tried, a 'total' and a 'military' one. Correspondingly, the deflator for country size is gross national product in the first case, military expenditure in the second. Thus the variables are:

EMP/GNP (.27)* = ratio of total civilian employment of U.S. government located in the country to gross national product, number of people as percent of GNP in millions of dollars [9, 17, 18, 20].

EMP_m/MIL (.14) = ratio of total civilian employment of U.S. Department of Defense located in the country to military expenditure, number of people as percent of military expenditure in millions of dollars [17, 19, 20].

Each of these two variables was tried both as a substitute for, and as a supplement to, the foreign-aid variable (AID/GNP). In the former situation they worked, but not as well as the 'aid' variable. In the latter situation collinearity with the aid variable prevented their successful inclusion in the regressions. There exists a measurement problem with the employment variables similar to that with the variable BOMBS: data are available only for one point in time (mid-1964). With annual data, the variables might have worked as a better substitute for the aid variable, although the collinearity problem would have remained.

III. The results

Results of the regression analysis are shown in Table 15.1. There are two groups of equations, *A* and *B*, corresponding to the intensive (military expenditure) and extensive (economic development 1 and 2) measures of power, respectively. For each group the cross-section regression was estimated for each of the years in the sample period (1958–67). Because of various limitations in the data, the sample size (a maximum of 76 countries) differs over the moving cross-section regression; degrees of freedom are shown as *d.f.* Below each explanatory variable its regression coefficient and *t* value (in parentheses) are presented left to right. A constant term (not shown) was included, and significant, in all the regressions. The last column shows *R* squared corrected for degrees of freedom ($\bar{R}^2$).

To what extent do these regressions confirm the hypotheses presented at the end of section 1? Hypotheses 1*a* and 1*b* are satisfied if the regression coefficients of all explanatory variables have the expected signs. The susceptibility of a country to U.S. threatening and bribing actions is represented by TRADE/GNP and AID/GNP, respectively, and their effect on foreign-exchange/reserves is positive as expected (hypothesis 1*a*). The power of a country is measured by the variables MIL (Group A) and EC DEV 1 and 2 (Group B). The outer country with a greater power would be less susceptible to pressure to

Table 15.1 Regression equations

			Group A			
Year	*d.f.*	*TRADE/GNP*	*MIL*		*AID/GNP*	$\bar{R}^2$
1958	55	0.46 (0.97)	–.0099 (2.06)		1.18 (1.60)	.11
1959	53	0.77 (1.46)	–.0105 (2.14)		1.84 (1.78)	.16
1960	54	0.58 (1.11)	–.0085 (1.89)		2.03 (2.30)	.16
1961	58	0.74 (1.49)	–.0065 (1.52)		2.22 (1.90)	.13
1962	60	0.75 (1.72)	–.0070 (1.80)		1.01 (0.91)	.11
1963	62	0.92 (2.25)	–.0068 (2.06)		1.77 (1.63)	.19
1964	64	1.12 (3.77)	–.0061 (2.22)		3.45 (2.69)	.32
1965	65	0.99 (3.27)	–.0069 (2.60)		3.15 (2.85)	.32
1966	70	0.98 (3.36)	–.0066 (2.66)		3.29 (2.58)	.31
1967	65	0.98 (3.11)	–.0050 (2.07)		3.44 (1.90)	.25
			Group B			
Year	*d.f.*	*TRADE/GNP*	*EC DEV 1*	*EC DEV 2*	*AID/GNP*	$\bar{R}^2$
1958	57	1.10 (3.46)	–7.72 (2.32)	–14.64 (4.57)	0.53 (0.79)	.34
1959	53	1.17 (3.34)	–8.53 (2.34)	–12.00 (3.59)	1.17 (1.22)	.32
1960	54	1.17 (3.69)	–7.37 (2.17)	–13.20 (4.41)	1.73 (2.20)	.40
1961	56	1.23 (3.73)	–7.31 (2.00)	–12.16 (3.99)	1.95 (1.67)	.34
1962	55	1.47 (4.18)	–6.90 (1.70)	–12.02 (3.69)	0.74 (0.66)	.32
1963	57	1.58 (5.02)	–7.23 (2.20)	–11.04 (3.82)	1.66 (1.63)	.40
1964	56	1.50 (5.29)	–3.56 (1.14)	–6.57 (2.43)	4.08 (3.16)	.42
1965	56	1.36 (4.44)	–5.27 (1.60)	–6.82 (2.37)	3.33 (2.79)	.37
1966	59	1.42 (4.80)	–6.57 (2.27)	–6.76 (2.53)	3.38 (2.58)	.38
1967	56	1.20 (4.13)	–5.17 (1.74)	–6.48 (2.42)	3.59 (1.90)	.32

hold a higher foreign-exchange/reserves. ratio, implying a negative regression coefficient for all power variables. The regressions have this property, thus satisfying hypothesis 1*b*. Therefore hypotheses 1*a* and 1*b* are confirmed.

Hypotheses 2*a* and 2*b* are tested by examining the movements of individual regression coefficients over time. Increased application of U.S. pressure as the system progresses into the crisis zone would be reflected in higher coefficients for TRADE/GNP and AID/GNP (hypothesis 2*a*); increased restraint on the part of the outer countries would appear in the form of a lower absolute value of the coefficients of MIL, EC DEV 1, and EC DEV 2 (hypothesis 2*b*). Thus for hypotheses 2*a* and 2*b* to be confirmed, all regression coefficients would have to exhibit an upward movement in their algebraic value over time. The results show that no variable exhibits a persistent positive trend in its regression coefficient over time. Therefore hypotheses 2*a* and 2*b* are not confirmed.

However, a weakened form of the hypotheses can be accepted. All regression coefficients are higher at the end than at the beginning of the sample

Table 15.2 Net impact of restraining forces

Year	*Group A*	*Group B*
1958	.07	−.30
1959	.15	−.17
1960	.19	−.08
1961	.22	−.07
1962	.10	−.10
1963	.20	.06
1964	.44	.50
1965	.35	.35
1966	.32	.26
1967	.31	.20

period. For all variables in each group of regressions, the size of the coefficient in 1965–67 is uniformly higher than in 1958–60. Now, in the majority of cases (AID/GNP in both Groups, EC DEV 1, and EC DEV 2) the change occurred primarily because of a large jump in the coefficient between 1963 and 1964. For the other variables (TRADE/GNP and MIL) this property does not hold. However, the 1963–64 jump dominates the equations over time, as shown by an analysis of beta coefficients.

Suppose one were to add the beta coefficients of the factors making for restraint (TRADE/GNP and AID/GNP) and from this sum subtract that of the beta coefficients of the power variables, representing the factors which enable a country to resist the forces that induce restraint in foreign-exchange/reserves behavior. The result would represent the net impact of the restraining forces, and is shown in Table 15.2.

For both Groups A and B the net impact of restraining forces in 1965–67 is again uniformly greater than in 1958–60, and furthermore the bulk of this increase in restraint can be accounted for by the movement between 1963 and 1964. There is the suggestion of two periods at work, 1958–63 and 1964–67, with restraint in foreign-exchange/reserves behavior greater in the latter period. Table 15.2 exhibits the further intriguing result that net restraint reached a peak in 1964 and declined thereafter, a behavior quite inconsistent with the Officer–Willett analysis of behavior in the crisis zone.

One is left with two events to be explained. (1) What happened in 1963–64 that caused a significant shift to more restraint in the foreign-exchange/reserves behavior of the outer countries; or, expressed differently, why did the restraint occur largely in a discontinuous rather than continuous fashion? (2) Why did this restraint weaken after its peak in 1964?

These questions cannot be answered with certainty; one can offer, however, an explanation that seems reasonable. The answer lies in expanding the Officer-Willett model to include the interrelationship of foreign-exchange/reserves restraint with negotiations to reform the international monetary system. Restraint reached a peak in 1964, in response to the U.S. initiative in organizing meetings of the Group of Ten in 1963–64 with the aim of negotiating a reform of the system. The jump in the outer countries' restraint in 1963–64 represented a 'bribe' to the United States for accepting the principle of changing the international monetary system via negotiations. This would be the explanation of the discontinuous upward shift in restraint in 1963–64.

Now, it became apparent at an early stage of the talks that the United States was determined to veto any proposed reform of the system that threatened the preeminent role of the dollar.[12] The outer countries responded to this position by weakening restraint in their foreign-exchange/reserves behavior. While France was the only country to engage in large-scale gold conversions beginning in early 1965, this does not imply that the French behavior alone is the explanation of the 1964 peak in the outer countries' restraint. Rather, the shift in French policy was indicative of a general reduction in restraint, a reduction that was measurable but not large except in the case of France. One major country alone could exchange significant amounts of dollars for gold without causing a collapse of the system, and France preempted this position. The remaining participants reduced their restraint, as measured by their foreign-exchange/reserves ratio, but not sufficiently to bring down the system. Their goal was to put pressure on the United States to take a more flexible position in the negotiations to reform the international monetary system, negotiations which culminated in the adoption of Special Drawing Rights.

IV. Power relationships versus portfolio management as an explanation of reserve-asset behavior

In this Chapter I have developed and tested a theory of reserve-asset behavior in the crisis zone that is based on power relationships. This approach contrasts with the traditional analysis of reserve-asset composition, which involves portfolio balancing of the risks and returns of assets and is represented best by the model of Peter B. Kenen [11]. Kenen viewed countries' reserve-asset behavior as solely a matter of portfolio management: 'their choice as between gold and dollars is governed by the interest reward that dollar balances offer and by the foreigners' estimate of the risk that the dollar price of gold may rise. This risk, in turn, will be assumed to vary inversely with the American reserve ratio – the ratio of American gold to short-term liabilities (foreign-owned dollar balances)' [11, pp. 572–73]. Thus, on the part of the outer countries, there is no concern at all with the preservation of the international monetary system.[13]

Kenen himself was the first to provide a test of the portfolio-management theory of reserve assets [12]. Using quarterly observations over the time period 1950–60, he investigated the demand for gold on the part of some 61 countries. His basic regression equation is the change in gold holdings as a function of the change in total reserves, a trend term, and foreign-exchange holdings lagged one quarter. The model simply involves a proportional relationship between first-differenced gold and reserves, corrected, first, by a trend term to measure changes in gold versus foreign-exchange preferences over time and, second, by a variable (formally the difference between actual and desired foreign-exchange holdings)[14] to reflect lags in portfolio adjustment. Even taking the portfolio-management approach as given, Kenen's empirical work searches only for mechanistic behavior on the part of central banks – no variables serve directly to measure the return and risk of holding gold relative to dollars.

Margaret L. Greene [7] redid Kenen's work for a later time period (1957–64), confining herself to the ten most important outer countries – the Group of Ten (excluding the United States) and Switzerland. Greene is aware of the mechanistic nature of Kenen's approach; she writes: 'In brief no attempt was made to explain central-bank reserve-asset preferences in terms of economic arguments such as the costs and risks of holding a particular kind of reserve asset' [7, p. 74]. Greene decided to expand Kenen's approach by introducing into the regressions variables which would affect portfolio management. In particular: (1) Kenen's estimate of working balances as the mean foreign-exchange holdings over the period is replaced by a sophisticated measure based on a moving standard deviation of the change in total reserves. (2) The U.S. treasury-bill rate is tried as 'a proxy for interest earnings on dollar balances . . . for the opportunity cost of holding gold' [7, pp. 377, 378]. (3) A stock-adjustment approach is utilized whereby the change in gold holdings is proportional to the difference between desired and (lagged) actual holdings. (4) Several alternative proxies to measure the risk of a devaluation of the dollar are tried – the U.S. deficit, U.S. gold loss, U.S. reserve ratio, and the change in that ratio.

Thus Greene carries Kenen's model to its logical culmination as far as data availability permits. She adopts the Tobin-Markowitz theory, its applicability to reserve-asset preferences involving the assumptions that 'central banks are risk-averting utility maximizers' and that 'the central bank's utility is assumed to vary directly with the expected return and inversely with its variance for the entire portfolio' [7, p. 381]. It turns out that, as Greene herself admits, 'the regression equations are unsatisfactory. . . . The income and risk proxies were not well-behaved, the interest-rate coefficients are negative more often than positive, and the risk proxies are seldom significant' [7, p. 382]. So Greene shows that, when subjected to sophisticated testing, the portfolio-balance approach fails as a theory which explains reserve-asset preferences.

Helmut A. Hagemann has provided a paper [8] that he presents as an extension and testing of the Kenen model. The main part of his study is unlike the work of Kenen and Greene (but similar to mine) in the type of sample (cross-section rather than time series) and the dependent variable (foreign-exchange / reserves). The one point in time examined is the year 1964. Hagemann argues against a power-political determination of reserve-asset preferences, the keynote of my paper. He writes: 'The view that the portfolio policies of central banks are dominated by politics regardless of any economic considerations seems to be widespread' [8, p. 65]. He concludes that his results show 'there is strong evidence that the reserve policies of central banks are determined to a large extent by economic factors' [8, p. 76]. I cannot agree with Hagemann's interpretation of the conventional wisdom on the subject. Surely it is the Kenen view, the portfolio-management approach, which held sway for a long time.

The issue is more than a matter of semantics. Portfolio management involves, let us say, 'narrow' economic factors. Hagemann looks at 'broader' economic factors. I should call some of his factors 'political'; economic size would be an example. However, whether a factor is predominantly 'economic' or 'political' depends on the interpretation. Hagemann has a trade variable similar to mine, except that the flows are restricted to exports alone rather than exports plus imports. He writes: 'Moreover, the closer the trade ties with the United States, the more likely a country will be to follow a devaluation of the dollar and avoid embarassing capital losses on its balance sheet' [8, p. 65]. This is a ('narrow') economic interpretation of the trade variable. (Granted, it is not Hagemann's only interpretation.) I give the trade variable a power-political interpretation: it measures U.S. leverage over the country's economy.

Hagemann uses the *flow* of direct investment (relative to GNP) as an explanatory variable, and it has a significantly negative effect on foreign-exchange/reserves. 'Their [outer countries'] hidden or declared reluctance to finance direct U.S. investment by accumulating short-term dollars is reflected in their portfolio policies' [8, p. 69]. On the other hand, Hagemann finds that the ratio of the flow of portfolio investment to GNP comes in with a positive sign. 'This means that the official institutions have generally welcomed U.S. long-term capital, unless it is associated with a shift of control to U.S. investors' [8, p. 69].

Given the 'power' standpoint of my theory, the best such investment variable would be the ratio of the *stock* of U.S. direct investment to the recipient country's GNP.[15] A priori, this variable would have a *positive* effect on dollars/reserves as distinct from Hagemann's result of a negative impact of the corresponding flow variable. A country with substantial U.S. control of its economy presumably would hold a higher dollars/reserves ratio than otherwise. One notes that the investment variable is thus an indicator of economic dependence on the United States. However, my trade variable (TRADE/GNP) is a better measure. After all, direct investment is a two-edged sword, possibly leading to acrimony with the recipient country. More to the point, U.S. companies and subsidiaries abroad can be appropriated, and the plants remain. True, new investment might cease to flow, technicians might leave, etc.; but, alternatively, the United States and the companies concerned might decide to make the best of the situation. In any event, the physical plant can be captured easily. In the case of trade, though, there is only the flow to consider. The outer country has no leverage – stemming the flow can only harm itself.

Hagemann considers the effects of economic size (measured by GNP) and wealth (measured by GNP per capita) on a country's dollars/reserves ratio. One would expect a high, though not perfect, correlation between the two variables. A priori, Hagemann sees economic size as having a negative, wealth an uncertain, effect on dollars/reserves. However, it turns out that the regression coefficients of both GNP and per capita GNP are positive. This is at variance with my results, in which GNP and similar measures (military expenditure and economic-development variables) have a negative impact on the dollars/reserves ratio.

Furthermore, Hagemann and I differ in our *interpretation* of such variables. He writes: 'Thus, in this sample, the wealthy countries have adopted a higher D_i/R_i ratio than the poorer countries, because to the wealthy countries even large devaluation losses may be less harmful than a disruption of international trade and capital movements' [8, p. 70]. Hagemann is still hung up on portfolio management. He sees the most powerful countries as having *even greater* interest in portfolio management of their reserves than do the minor countries! The chief lesson of the Officer-Willett approach to the international monetary system [14, 15] and a keynote of the present study is that within the crisis zone (wherein Hagemann's time point, 1964, lies) one would expect precisely the opposite behavior.

Concerning the positive effect of economic size (GNP), Hagemann argues that the explanation is that a large country has a greater stake in the system, a greater impact on the system, and more publicity attached to its actions. However, Hagemann overlooks the desire for individual advantage that is coupled with concern for the system.[16] Furthermore, powerful countries are less susceptible to U.S. pressure to act to preserve the system, i.e., to hold a higher dollars/reserves ratio.

The problem, of course, is that Hagemann refuses to interpret his GNP variables as measures of power, analogous to the measures of military expenditure and economic development that I employ. All countries have a stake in the system, the less as well as the more powerful. While there is greater incentive on the powerful countries to act to preserve the system (as their impact on it is greater), they also have a greater opportunity to evade responsibility with impunity – to take a 'free ride' (in the form of a higher dollars/reserves ratio) as long as there remains some slack in the system. Nevertheless, Hagemann does move part way to my position. He complements his GNP variables with two alternative reserves measures, a country's reserves relative to total non-U.S. reserves, as one variable, and relative to GNP, as the other. It turns out that these variables have a *negative* effect on the dependent variable. This result corresponds to the negative impact of my power variables. However, with only a one-point cross-section, Hagemann cannot test for growing restraint over time.

Hagemann uses the same foreign-aid variable that I do (AID/GNP). However, even here he misinterprets the meaning of the variable. He interprets it as part of the total capital flow the country receives from the United States (direct, portfolio, and 'unrequited' investment). The concern is with U.S.

restrictive and deflationary policies, resulting from a fall in the U.S. reserve ratio and leading to a decline in the outer country's capital inflow. I interpret foreign aid in a more direct manner: if the U.S. objects to a country's dollars/reserves policy, it can threaten to reduce or eliminate aid. Hagemann, on the other hand, looks on aid as being just a source of foreign exchange, whereas in reality it is much more than that. Military aid involves weaponry hardware and technical assistance; economic aid includes the receipt of U.S. technology and equipment, and the possibility of future, continuing, or additional military aid. Again it is a matter of power, increments to the outer country's economic and military power, let us say 'in return for' restraint in its reserve-asset policy.

Hagemann recognizes the possibility of restraint to preserve the international monetary system. However, he sees it as a future possibility rather than an event which actually had occurred by the time of his study. 'We realize that past reserve policies may be revised drastically in the face of changed circumstances. Our equations are based on data from a period when the breakdown of the entire international monetary system was a remote possibility. Now, as this danger is growing, central banks may depart from past policies. Some may increase their D_i/R_i ratios to help forestall such a crisis, others may reduce their D_i/R_i ratios below the level we would expect from our equations to reduce the risk of suffering large capital losses' [8, p. 74]. Apparently, Hagemann is agnostic as to which effect would dominate. In contrast, my results show that it is the former (restraint to preserve the system) rather than the latter (avoidance of risk in portfolio return) which prevailed in the decade 1958–67. Indeed, the very year of Hagemann's cross-section analysis (1964) witnessed a significant shift to more restraint in the reserve-asset behavior of the outer countries.

Notes

* Originally published in *Journal of Money, Credit, and Banking*, vol. 6 no. 2 (May 1974), pp. 191–211. Incorporates 'Errata,' *Journal of Money, Credit, and Banking*, vol. 7, no. 3 (August 1975), p. 407. Reproduced with permission of The Ohio State University Press.

1 For critiques of both the Kenen and Mundell models of the international monetary system, see [14].

2 It might be objected that following the U.S. suspension of gold conversion in August, 1971, the private-market price of gold actually increased rather than decreased. However, that U.S. policy change in essence represented a safeguarding of its gold holdings rather than a repudiation of gold. The private gold price would have declined – and probably below $35 an ounce – if the United States had announced the opposite policy, namely, that 'gentlemen's agreements' limiting convertibility were no longer necessary, that the outer countries were permitted, indeed encouraged, to exchange dollars for gold until the U.S. gold ran out, and that thereafter the United States would no longer buy gold or have any desire to own it.

3 Canada and Paraguay.

4 There exist rare exceptions. For example, Nepal (not in the sample) holds about 20

percent of its foreign exchange in Indian rupees. In general, countries outside the sterling area that are in a dependent relationship to a more powerful country, emanating from a former colonial status, are excluded from the sample – though for reason of lack of data. Members of the 'franc area' would be included in this category.

5 Although the true numerator (D) is available for Canada and Paraguay, the information was ignored in order to have consistent series.

6 Definition of a variable is followed by reference to its data source. Where necessary, conversion of the denomination of a series from domestic currency to U.S. dollars was performed using exchange-rate information found in [9] and [18].

7 This procedure differs from the usual one of imposing a zero mean and unit standard deviation on the variables. Russett presents a convincing argument that the conventional technique would be inappropriate for his data [16, p. 22, fn. 5].

8 Russett also derived a set of principal components obtained from only 29 of the underlying variables, those highly correlated (.60 or greater) with one of the four factors (excluding 'size'). These principal components were tried as an alternative set of explanatory variables, but they were less significant than those derived from all 54 original variables.

9 Yugoslavia has the closest to a 'communist' government of the 76 countries.

10 Correspondingly, in the experimental regressions involving total-dollars/reserves (D_t/R) as the dependent variable, trade with the United Kingdom was excluded from the trade variables.

11 Hagemann states that by accepting a low gold content of its reserves, the outer country 'can help the reserve-currency countries to avoid harsh balance-of-payments measures which would be most harmful to its own foreign trade' [8, p. 65].

12 See, for example, the discussion in Cohen [4, chs. 2–3].

13 For a critique of this aspect of the Kenen model, see [14, pp. 689–92].

14 Desired foreign exchange does not enter the equation directly. Structural coefficients are inferred using the mean foreign-exchange holdings over the sample period as an approximation to desired foreign exchange. As Greene has pointed out, 'His procedure assumes that working-balance requirements are constant throughout the period and equal to the mean of *all* foreign-exchange holdings' [7, p. 375]. Such an assumption is surely not tenable.

15 Data limitations (too few countries for which estimates of U.S. direct investment are published) prevented my use of this variable.

16 In this respect he is committing what may be called '[Harry G.] Johnson's error.' See [15, p. 58, fn. 18].

References

[1] Agency for International Development. *U.S. Overseas Loans and Grants and Assistance from International Organizations, Obligations and Loan Authorizations*.

[2] Alger, C. F. and S. J. Brams. 'Patterns of Representation in National Capitals and Intergovernmental Organizations.' *World Politics*, **19** (July 1967), 646–63.

[3] Beaton, L. *Must the Bomb Spread?* Middlesex, England: Penguin Books Ltd., 1966.

[4] Cohen, S. D. *International Monetary Reform, 1964–69*. New York: Praeger Publishers, 1970.

[5] Deutsch, K. W. *The Analysis of International Relations*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1968.

[6] Federal Reserve Bank of St. Louis. *U.S. Balance of Payments Trends*.

[7] Greene, M. L. 'Reserve-Asset Preferences Revisited.' In P. B. Kenen and R. Lawrence, eds., *The Open Economy: Essays on International Trade and Finance*. New York: Columbia University Press, 1968. Pp. 355–85.

[8] Hagemann, H. A. 'Reserve Policies of Central Banks and their Implications for U.S. Balance of Payments Policy.' *American Economic Review*, **59** (March 1969), 62–77.

[9] International Monetary Fund. *International Financial Statistics*.

[10] International Monetary Fund and International Bank for Reconstruction and Development. *Direction of Trade*.

[11] Kenen, P. B. 'International Liquidity and the Balance of Payments of a Reserve-Currency Country.' *Quarterly Journal of Economics*, **74** (November 1960), 572–86.

[12] —— *Reserve-Asset Preferences of Central Banks and Stability of the Gold Exchange Standard*. Princeton Studies in International Finance, No. 10. Princeton, N.J.: Princeton University Press, 1963.

[13] Mundell, R. A. 'The Crisis Problem.' in *International Economics*. New York: The Macmillan Company, 1968. Pp. 282–97.

[14] Officer, L. H. and T. D. Willett. 'Reserve-Asset Preferences and the Confidence Problem in the Crisis Zone.' *Quarterly Journal of Economics*, **83** (November 1969), 688–95.

[15] —— 'The Interaction of Adjustment and Gold-Conversion Policies in a Reserve-Currency System.' *Western Economic Journal*, **8** (March 1970), 47–60.

[16] Russett, B. M. *International Regions and the International System: A Study in Political Ecology*. Chicago: Rand McNally & Company, 1967.

[17] Sapin, B. M. *The Making of United States Foreign Policy*. Washington, D.C.: The Brookings Institution, 1966.

[18] Statistical Office of the United Nations. *Yearbook of National Accounts Statistics*.

[19] Stockholm International Peace Research Institute. *SIPRI Yearbook of World Armaments and Disarmament*. Stockholm: Almqvist & Wiksell, 1969.

[20] U.S. Arms Control and Disarmament Agency. *World Military Expenditures 1969*.

[21] U.S. Department of Commerce, Office of Business Economics. *Survey of Current Business*.

16 Afterword to Part IV

Fascination with gold

Gold has always been an object of fascination to both scholars and the public at large, partly because of the use of gold as coined money, in jewelry, and in the arts, partly because of gold's high value/weight ratio. The high value/weight ratio can also be described as a high price of gold per physical unit. Gold is usually measured in U.S. dollars per ounce or grain. The measurement system for gold is the Troy system, in which 1 Troy ounce equals 480 grains and 12 Troy ounces equals 1 Troy pound. The Troy system is used worldwide not only for gold but also for other precious metals and gems. Under the avoirdupois system of measurement, prevalent domestically in the United States (but not in the rest of the world or in international trade, where the metric system reigns), 1 avoirdupois ounce = 437.5 grains, and 1 avoirdupois pound = 16 avoirdupois ounces. Note that the grain is the same unit under both systems. So the price of gold expressed as dollars per grain is unambiguous; but expressed as dollars per ounce, the ounce is almost ten percent (480/437.5) heavier than the ounce unit commonly understood in the United States.

Gold is, and has been, a focus in literature, the arts, plays and film. A few examples follow. Gold is mentioned early in the Bible: 'there is gold; and the gold of that land is good' (Genesis 2), and Kettell (1982: 19) observes, with this citation, that gold is the first element so mentioned. Shakespeare writes: 'All that glisters is not gold' (*The Merchant of Venice*, scene 7). There is the James Bond film (and novel) *Goldfinger*. Historically, gold at times has been of tremendous importance. Under the gold standard, which finally ended in the 1930s, a country's monetary unit was defined in terms of gold and all other forms of money were exchangeable in terms of gold.

The post-World War II gold-exchange standard had several attributes of the old gold standard. In particular, gold and U.S. dollars were interchangeable within a narrow band around the U.S.-Treasury fixed price of $35 an ounce, and exchange rates between currencies were also fixed. During that period, some foresaw a worthless gold in the future. There is the famous statement of Premier Nikita S. Khruschev of the Soviet Union: 'Lenin said the day would come when gold would serve to coat the walls and floors of public toilets' – quoted in Triffin (1960: 195).

However, that prediction did not come to pass, even after the gold-exchange standard was dismantled, and in several steps. In 1968, the private and official markets for gold were decoupled, and the private price of gold was no longer hinged to the official price of $35 an ounce. In 1971, officially held dollars were formally denied convertibility into gold. In 1973, the fixed exchange-rate system was replaced by floating exchange rates.

Today, there is a thriving private market in gold, as gold is a traded commodity in various markets throughout the world. The motivation is speculation on the price of gold, or investment in gold as an asset the long-term price of which might be forecasted as upward. Further, in some developing countries, gold is (as it historically has been) a basic store of value – hoarded by people who distrust government paper money (currency), bank money (deposits), and paper assets (bonds).

Documentary histories of gold

Chapter 14 is a review article of a book edited by Mark Duckenfield and consisting of collected historical documents on gold. This article is presented here, because it compares the Duckenfield work with six other volumes that have the same theme. All seven volumes are of potential use to the scholar interested in the history of gold. The volumes have varying strengths and weaknesses, which sometimes reinforce and sometimes complement each other. Rare for a book review or review article, chapter 14 includes tables, six in all, that facilitate understanding of the virtues and limitations of the seven works considered.

Central-bank reserve-asset preferences

Chapter 15 investigates a particular period in the history of the gold-exchange standard. Between 1958 and 1967, gold and U.S. dollars were the primary reserve assets of central banks. Neither SDRs nor the euro existed. The chapter studies the *composition* of outer-country (meaning non-U.S.) central-bank reserves, as distinct from the total demand for such reserves. Reserve-asset composition means the distribution of total reserves among components; in 1958–1967 this distribution was principally between dollars and gold.

A survey of the literature on central-bank demand for reserves is in International Monetary Fund (2000b: 14–26). A review of the literature on the composition of reserves, but written prior to the original publication of Chapter 15 (the year 1974) is Williamson (1973), with theoretical and empirical work discussed on pages 697–701 and 701–703. This empirical work is also reviewed in Chapter 15 (Section IV).

The theme of Chapter 15 is that in the 'crisis zone,' which existed in

1958–1967, a political – in particular, 'power' – motive for the determination of reserve-asset composition existed, and this approach is found to work empirically. Emphatically, as argued in the chapter, the portfolio-management motive was unimportant. Nevertheless, empirical studies of reserve-asset composition that incorporate the time period 1958–1967 adopt the portfolio-management approach. In addition to the studies of this nature reviewed in Chapter 15 and in Williamson (1973), Masera and Falchi (1982) and Bernholz (2002) fall into this category. The latter works were written after Chapter 15 first appeared (1974); so it is disheartening that the authors make no reference to the 'power' approach.

On the other hand, portfolio-management studies of reserve-asset composition in periods after 1958–1967 are legitimate. Such studies are even more defensible, if they are confined to the currency composition of the given total of the foreign-exchange component of reserves; thus the gold-versus-dollar (or, gold versus foreign-exchange) decision is considered predetermined. These studies include Heller and Knight (1978), Ben-Basat (1984), and Dooley, Lizondo and Mathieson (1989).

References

Ben-Basat, A. (1984) *Reserve-Currency Diversification and the Substitution Account*. Princeton: International Finance Section, Princeton University.

Bernholz, P. (2002) 'Advantages and Disadvantages of the Holding of Gold Reserves by Central Banks – With Special Reference to the Swiss National Bank', *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* **138**: 99–113.

Dooley, M. P., J. S. Lizondo, and D. J. Mathieson (1989) 'The Currency Composition of Foreign Exchange Reserves', IMF *Staff Papers*, **36**: 385–434.

Heller, H. R. and M. Knight (1978) *Reserve-Currency Preferences of Central Banks*. Princeton: International Finance Section, Princeton University.

Kettell, B. (1982) *Gold*. Cambridge, MA: Ballinger.

Masera, R. S. and G. Falchi (1982) 'The Role of Gold in Central Banks' Reserve Composition', in A. Quadrio-Curzio (ed.) *The Gold Problem: Economic Perspectives*, Oxford: Oxford University Press.

Triffin, R. (1960) *Gold and the Dollar Crisis*. New Haven: Yale University Press.

Williamson, J. (1973) 'Surveys in Applied Economics: International Liquidity', *Economic Journal*, **83**: 685–746.

Index of names

Index of subjects