

MILTON FRIEDMAN ON ECONOMICS



SELECTED PAPERS

MILTON FRIEDMAN

ON ECONOMICS

SELECTED PAPERS

MILTON FRIEDMAN

ON ECONOMICS

SELECTED PAPERS

WITH AN AFTERWORD BY GARY S. BECKER

THE UNIVERSITY OF CHICAGO PRESS
CHICAGO AND LONDON

The essays in the volume originally appeared in issues of *The Journal of Political Economy* and *The Journal of Law and Economics*.

The University of Chicago Press, Chicago 60637
The University of Chicago Press, Ltd., London

11 10 09 08 07 5 4 3 2 1

© 2007 by The University of Chicago
All rights reserved. Published 2007
Printed in the United States of America

The 1976 Nobel Prize in Economic Sciences Lecture “Inflation and Unemployment” is © The Nobel Foundation 1976.

Library of Congress Cataloging-in-Publication Data

Friedman, Milton, 1912–2006.

Milton Friedman on economics : selected papers / Milton Friedman
p. cm.

ISBN 978-0-226-26349-6 (pbk. : alk. paper)

1. Economics. 2. Monetary policy. I. Savage, Leonard J. II. Becker, Gary Stanley, 1930– III. Title. IV. Title: On economics.

HB171.F776 2007

330.15'53092—dc22

2007031094

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984. ☺

MILTON FRIEDMAN ON ECONOMICS

- 1 Nobel Lecture: Inflation and Unemployment
Milton Friedman
The Journal of Political Economy, Vol. 85, No. 3 (Jun., 1977),
pp. 451–472
- 23 The Utility Analysis of Choices Involving Risk
Milton Friedman and L. J. Savage
The Journal of Political Economy, Vol. 56, No. 4 (Aug., 1948),
pp. 279–304
- 57 The Expected-Utility Hypothesis and the Measurability of
Utility
Milton Friedman and L. J. Savage
The Journal of Political Economy, Vol. 60, No. 6 (Dec., 1952),
pp. 463–474
- 72 A Statistical Illusion in Judging Keynesian Models
Milton Friedman and Gary S. Becker
The Journal of Political Economy, Vol. 65, No. 1 (Feb., 1957),
pp. 64–75
- 89 The Demand for Money: Some Theoretical and Empirical
Results
Milton Friedman
The Journal of Political Economy, Vol. 67, No. 4 (Aug., 1959),
pp. 327–351
- 120 Interest Rates and the Demand for Money
Milton Friedman
The Journal of Law and Economics, Vol. 9 (Oct., 1966), pp. 71–
85
- 135 Government Revenue from Inflation
Milton Friedman
The Journal of Political Economy, Vol. 79, No. 4 (Aug., 1948),
pp. 846–856
- 146 The Crime of 1873
Milton Friedman
The Journal of Political Economy, Vol. 98, No. 6 (Dec., 1990),
pp. 1159–1194
- 181 Afterword
Milton Friedman as a Microeconomist
Gary S. Becker
- 187 Index

Nobel Lecture: Inflation and Unemployment

When the Bank of Sweden established the prize for economic science in memory of Alfred Nobel in 1968, there doubtless was—as there doubtless still remains—widespread skepticism among both scientists and the broader public about the appropriateness of treating economics as parallel to physics, chemistry, and medicine. These are regarded as “exact sciences” in which objective, cumulative, definitive knowledge is possible. Economics and its fellow social sciences are regarded more nearly as branches of philosophy than of science properly defined, enmeshed with values at the outset because they deal with human behavior. Do not the social sciences, in which scholars are analyzing the behavior of themselves and their fellowmen, who are in turn observing and reacting to what the scholars say, require fundamentally different methods of investigation than the physical and biological sciences? Should they not be judged by different criteria?

I. Social and Natural Sciences

I have never myself accepted this view. I believe that it reflects a misunderstanding not so much of the character and possibilities of social science as of the character and possibilities of natural science. In both, there is no “certain” substantive knowledge; only tentative hypotheses that can never be “proved” but can only fail to be rejected, hypotheses in which we may have more or less confidence, depending on such features as the breadth of experience they encompass relative to their

I am much indebted for helpful comments on the first draft of this paper to Gary Becker, Karl Brunner, Phillip Cagan, Robert Gordon, Arnold Harberger, Harry G. Johnson, S. Y. Lee, James Lothian, Robert E. Lucas, David Meiselman, Allan Meltzer, José Scheinkman, Theodore W. Schultz, Anna J. Schwartz, Larry Sjaastad, George J. Stigler, Sven-Ivan Sundqvist, and the participants in the Money and Banking Workshop of the University of Chicago. I am deeply indebted also to my wife, Rose Director Friedman, who took part in every stage of the preparation of the paper, and to my secretarial assistant, Gloria Valentine, for performance above and beyond the call of duty.

own complexity and relative to alternative hypotheses, and the number of occasions on which they have escaped possible rejection. In both social and natural sciences, the body of positive knowledge grows by the failure of a tentative hypothesis to predict phenomena that the hypothesis professes to explain; by the patching up of that hypothesis until someone suggests a new hypothesis that more elegantly or simply embodies the troublesome phenomena, and so on ad infinitum. In both, experiment is sometimes possible, sometimes not (witness meteorology). In both, no experiment is ever completely controlled, and experience often offers evidence that is the equivalent of controlled experiment. In both, there is no way to have a self-contained closed system or to avoid interaction between the observer and the observed. The Gödel theorem in mathematics, the Heisenberg uncertainty principle in physics, the self-fulfilling or self-defeating prophecy in the social sciences all exemplify these limitations.

Of course, the different sciences deal with different subject matter, have different bodies of evidence to draw on (for example, introspection is a more important source of evidence for social than for natural sciences), find different techniques of analysis most useful, and have achieved differential success in predicting the phenomena they are studying. But such differences are as great among, say, physics, biology, medicine, and meteorology as between any of them and economics.

Even the difficult problem of separating value judgments from scientific judgments is not unique to the social sciences. I well recall a dinner at a Cambridge University college when I was sitting between a fellow economist and R. A. Fisher, the great mathematical statistician and geneticist. My fellow economist told me about a student he had been tutoring on labor economics, who, in connection with an analysis of the effect of trade unions, remarked, "Well surely, Mr. X (another economist of a different political persuasion) would not agree with that." My colleague regarded this experience as a terrible indictment of economics because it illustrated the impossibility of a value-free positive economic science. I turned to Sir Ronald and asked whether such an experience was indeed unique to social science. His answer was an impassioned no, and he proceeded to tell one story after another about how accurately he could infer views in genetics from political views.

One of my great teachers, Wesley C. Mitchell, impressed on me the basic reason why scholars have every incentive to pursue a value-free science, whatever their values and however strongly they may wish to spread and promote them. In order to recommend a course of action to achieve an objective, we must first know whether that course of action will in fact promote the objective. Positive scientific knowledge that enables us to predict the consequences of a possible course of action is clearly a prerequisite for the normative judgment whether that course

of action is desirable. The Road to Hell is paved with good intentions, precisely because of the neglect of this rather obvious point.

This point is particularly important in economics. Many countries around the world are today experiencing socially destructive inflation, abnormally high unemployment, misuse of economic resources, and, in some cases, the suppression of human freedom not because evil men deliberately sought to achieve these results, nor because of differences in values among their citizens, but because of erroneous judgments about the consequences of government measures: errors that at least in principle are capable of being corrected by the progress of positive economic science.

Rather than pursue these ideas in the abstract (I have discussed the methodological issues more fully in Friedman [1953]), I shall illustrate the positive scientific character of economics by discussing a particular economic issue that has been a major concern of the economics profession throughout the postwar period, namely, the relation between inflation and unemployment. This issue is an admirable illustration because it has been a controversial political issue throughout the period, yet the drastic change that has occurred in accepted professional views was produced primarily by the scientific response to experience that contradicted a tentatively accepted hypothesis—precisely the classical process for the revision of a scientific hypothesis.

I cannot give here an exhaustive survey of the work that has been done on this issue or of the evidence that has led to the revision of the hypothesis. I shall be able only to skim the surface in the hope of conveying the flavor of that work and that evidence and of indicating the major items requiring further investigation.

Professional controversy about the relation between inflation and unemployment has been intertwined with controversy about the relative role of monetary, fiscal, and other factors in influencing aggregate demand. One issue deals with how a change in aggregate nominal demand, however produced, works itself out through changes in employment and price levels; the other, with the factors accounting for the change in aggregate nominal demand.

The two issues are closely related. The effects of a change in aggregate nominal demand on employment and price levels may not be independent of the source of the change, and conversely the effect of monetary, fiscal, or other forces on aggregate nominal demand may depend on how employment and price levels react. A full analysis will clearly have to treat the two issues jointly. Yet there is a considerable measure of independence between them. To a first approximation, the effects on employment and price levels may depend only on the magnitude of the change in aggregate nominal demand, not on its source. On both issues, professional opinion today is very different than it was just after

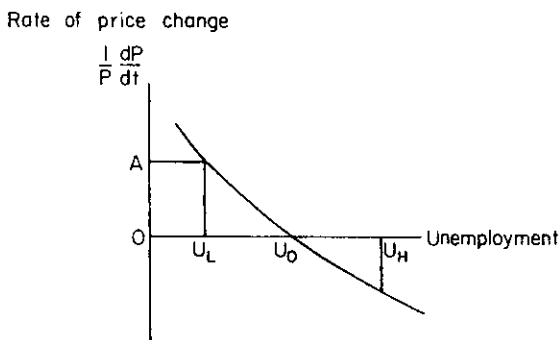


FIG. 1.—Simple Phillips curve

World War II because experience contradicted tentatively accepted hypotheses. Either issue could therefore serve to illustrate my main thesis. I have chosen to deal with only one in order to keep this lecture within reasonable bounds. I have chosen to make that one the relation between inflation and unemployment, because recent experience leaves me less satisfied with the adequacy of my earlier work on that issue than with the adequacy of my earlier work on the forces producing changes in aggregate nominal demand.

II. Stage 1: Negatively Sloping Phillips Curve

Professional analysis of the relation between inflation and unemployment has gone through two stages since the end of World War II and is now entering a third. The first stage was the acceptance of a hypothesis associated with the name of A. W. Phillips (1958) that there is a stable negative relation between the level of unemployment and the rate of change of wages—high levels of unemployment being accompanied by falling wages, low levels of unemployment by rising wages. The wage change in turn was linked to price change by allowing for the secular increase in productivity and treating the excess of price over wage cost as given by a roughly constant markup factor.

Figure 1 illustrates this hypothesis, where I have followed the standard practice of relating unemployment directly to price change, short-circuiting the intermediate step through wages.

This relation was widely interpreted as a causal relation that offered a stable trade-off to policymakers. They could choose a low unemployment target, such as U_L . In that case they would have to accept an inflation rate of A . There would remain the problem of choosing the measures (monetary, fiscal, perhaps other) that would produce the level of aggregate nominal demand required to achieve U_L , but if that were

done, there need be no concern about maintaining that combination of unemployment and inflation. Alternatively, the policymakers could choose a low inflation rate or even deflation as their target. In that case they would have to reconcile themselves to higher unemployment: U_0 for zero inflation, U_H for deflation.

Economists then busied themselves with trying to extract the relation depicted in figure 1 from evidence for different countries and periods, to eliminate the effect of extraneous disturbances, to clarify the relation between wage change and price change, and so on. In addition, they explored social gains and losses from inflation on the one hand and unemployment on the other, in order to facilitate the choice of the “right” trade-off.

Unfortunately for this hypothesis, additional evidence failed to conform with it. Empirical estimates of the Phillips curve relation were unsatisfactory. More important, the inflation rate that appeared to be consistent with a specified level of unemployment did not remain fixed: in the circumstances of the post-World War II period, when governments everywhere were seeking to promote “full employment,” it tended in any one country to rise over time and to vary sharply among countries. Looked at the other way, rates of inflation that had earlier been associated with low levels of unemployment were experienced along with high levels of unemployment. The phenomenon of simultaneous high inflation and high unemployment increasingly forced itself on public and professional notice, receiving the unlovely label of “stagflation.”

Some of us were skeptical from the outset about the validity of a stable Phillips curve, primarily on theoretical rather than empirical grounds (Friedman 1966*a*, 1966*b*, 1968*a*, 1968*b*). What mattered for employment, we argued, was not wages in dollars or pounds or kronor but real wages—what the wages would buy in goods and services. Low unemployment would, indeed, mean pressure for a higher real wage—but real wages could be higher even if nominal wages were lower, provided that prices were still lower. Similarly, high unemployment would, indeed, mean pressure for a lower real wage—but real wages could be lower, even if nominal wages were higher, provided prices were still higher.

There is no need to assume a stable Phillips curve in order to explain the apparent tendency for an acceleration of inflation to reduce unemployment. That can be explained by the impact of *unanticipated* changes in nominal demand on markets characterized by (implicit or explicit) long-term commitments with respect to both capital and labor. Long-term labor commitments can be explained by the cost of acquiring information by employers about employees and by employees about alternative employment opportunities plus the specific human capital that makes an employee’s value to a particular employer grow over time and exceed his value to other potential employers.

Only surprises matter. If everyone anticipated that prices would rise at, say, 20 percent a year, then this anticipation would be embodied in future wage (and other) contracts, real wages would then behave precisely as they would if everyone anticipated no price rise, and there would be no reason for the 20 percent rate of inflation to be associated with a different level of unemployment than a zero rate. An unanticipated change is very different, especially in the presence of long-term commitments—themselves partly a result of the imperfect knowledge whose effect they enhance and spread over time. Long-term commitments mean, first, that there is not instantaneous market clearing (as in markets for perishable foods) but only a lagged adjustment of both prices and quantity to changes in demand or supply (as in the house-rental market); second, that commitments entered into depend not only on current observable prices but also on the prices expected to prevail throughout the term of the commitment.

III. Stage 2: Natural Rate Hypothesis

Proceeding along these lines, we (in particular, E. S. Phelps [1967, 1970] and myself [1968*b*]) developed an alternative hypothesis that distinguished between the short-run and long-run effects of unanticipated changes in aggregate nominal demand. Start from some initial stable position and let there be, for example, an unanticipated acceleration of aggregate nominal demand. This will come to each producer as an unexpectedly favorable demand for his product. In an environment in which changes are always occurring in the relative demand for different goods, he will not know whether this change is special to him or pervasive. It will be rational for him to interpret it as at least partly special and to react to it by seeking to produce more to sell at what he now perceives to be a higher than expected market price for future output. He will be willing to pay higher nominal wages than he had been willing to pay before in order to attract additional workers. The real wage that matters to him is the wage in terms of the price of his product, and he perceives that price as higher than before. A higher nominal wage can therefore mean a lower *real* wage as perceived by him.

To workers, the situation is different: what matters to them is the purchasing power of wages not over the particular good they produce but over all goods in general. Both they and their employers are likely to adjust more slowly their perception of prices in general—because it is more costly to acquire information about that—than their perception of the price of the particular good they produce. As a result, a rise in nominal wages may be perceived by workers as a rise in real wages and hence call forth an increased supply at the same time that it is perceived by employers as a fall in real wages and hence calls forth an increased

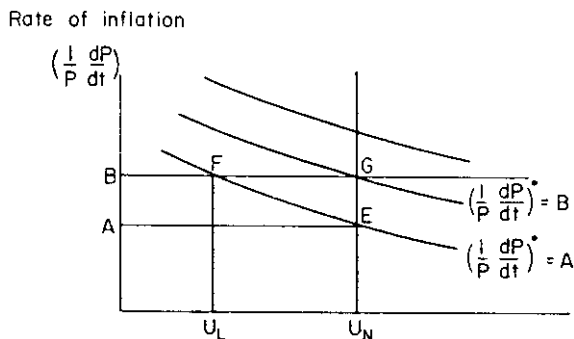


FIG. 2.—Expectations-adjusted Phillips curve

offer of jobs. Expressed in terms of the average of perceived future prices, real wages are lower; in terms of the perceived future average price, real wages are higher.

But this situation is temporary: let the higher rate of growth of aggregate nominal demand and of prices continue, and perceptions will adjust to reality. When they do, the initial effect will disappear and then even be reversed for a time as workers and employers find themselves locked into inappropriate contracts. Ultimately, employment will be back at the level that prevailed before the assumed unanticipated acceleration in aggregate nominal demand.

This alternative hypothesis is depicted in figure 2. Each negatively sloping curve is a Phillips curve like that in figure 1 except that it is for a particular anticipated or perceived rate of inflation, defined as the perceived average rate of price change, *not* the average of perceived rates of individual price change (the order of the curves would be reversed for the second concept). Start from point *E* and let the rate of inflation for whatever reason move from *A* to *B* and stay there. Unemployment would initially decline to U_L at point *F*, moving along the curve defined for an anticipated rate of inflation $[(1/P)(dP/dt)]^*$ of *A*. As anticipations adjusted, the short-run curve would move upward, ultimately to the curve defined for an anticipated inflation rate of *B*. Concurrently unemployment would move gradually over from *F* to *G*. (For a fuller discussion, see Friedman [1976], chap. 12.)

This analysis is, of course, oversimplified. It supposes a single unanticipated change whereas, of course, there is a continuing stream of unanticipated changes; it does not deal explicitly with lags, or with overshooting, or with the process of formation of anticipations. But it does highlight the key points: what matters is not inflation per se but unanticipated inflation; there is no stable trade-off between inflation and unemployment; there is a "natural rate of unemployment" (U_N)

which is consistent with the real forces and with accurate perceptions; unemployment can be kept below that level only by an accelerating inflation; or above it only by accelerating deflation.

The “natural rate of unemployment,” a term I introduced to parallel Knut Wicksell’s “natural rate of interest,” is not a numerical constant but depends on “real” as opposed to monetary factors—the effectiveness of the labor market, the extent of competition or monopoly, the barriers or encouragements to working in various occupations, and so on.

For example, the natural rate has clearly been rising in the United States for two major reasons. First, women, teenagers, and part-time workers have been constituting a growing fraction of the labor force. These groups are more mobile in employment than other workers, entering and leaving the labor market, shifting more frequently between jobs. As a result, they tend to experience higher average rates of unemployment. Second, unemployment insurance and other forms of assistance to unemployed persons have been made available to more categories of workers and have become more generous in duration and amount. Workers who lose their jobs are under less pressure to look for other work, will tend to wait longer in the hope, generally fulfilled, of being recalled to their former employment, and can be more selective in the alternatives they consider. Further, the availability of unemployment insurance makes it more attractive to enter the labor force in the first place, and so may itself have stimulated both the growth that has occurred in the labor force as a percentage of the population and its changing composition.

The determinants of the natural rate of unemployment deserve much fuller analysis for both the United States and other countries. So also do the meaning of the recorded unemployment figures and the relation between the recorded figures and the natural rate. These issues are all of the utmost importance for public policy. However, they are side issues for my present limited purpose.

The connection between the state of employment and the level of efficiency or productivity of an economy is another topic that is of fundamental importance for public policy but is a side issue for my present purpose. There is a tendency to take it for granted that a high level of recorded unemployment is evidence of inefficient use of resources, and conversely. This view is seriously in error. A low level of unemployment may be a sign of a forced-draft economy that is using its resources inefficiently and is inducing workers to sacrifice leisure for goods that they value less highly than the leisure under the mistaken belief that their real wages will be higher than they prove to be. Or a low natural rate of unemployment may reflect institutional arrangements that inhibit change. A highly static rigid economy may have a fixed place for everyone whereas a dynamic, highly progressive economy,

which offers ever-changing opportunities and fosters flexibility, may have a high natural rate of unemployment. To illustrate how the same rate may correspond to very different conditions: both Japan and the United Kingdom had low average rates of unemployment from, say, 1950 to 1970, but Japan experienced rapid growth, the United Kingdom, stagnation.

The “natural-rate” or “accelerationist” or “expectations-adjusted Phillips curve” hypothesis—as it has been variously designated—is by now widely accepted by economists, though by no means universally. A few still cling to the original Phillips curve; more recognize the difference between short-run and long-run curves but regard even the long-run curve as negatively sloped, though more steeply so than the short-run curves; some substitute a stable relation between the acceleration of inflation and unemployment for a stable relation between inflation and unemployment—aware of but not concerned about the possibility that the same logic that drove them to a second derivative will drive them to ever higher derivatives.

Much current economic research is devoted to exploring various aspects of this second stage—the dynamics of the process, the formation of expectations, and the kind of systematic policy, if any, that can have a predictable effect on real magnitudes. We can expect rapid progress on these issues. (Special mention should be made of the work on “rational expectations,” especially the seminal contributions of John Muth, Robert Lucas, and Thomas Sargent; see Muth [1961], Gordon [1976].)

IV. Stage 3: A Positively Sloped Phillips Curve?

Although the second stage is far from having been fully explored, let alone fully absorbed into the economic literature, the course of events is already producing a move to a third stage. In recent years higher inflation has often been accompanied by higher, not lower, unemployment, especially for periods of several years in length. A simple statistical Phillips curve for such periods seems to be positively sloped, not vertical. The third stage is directed at accommodating this apparent empirical phenomenon. To do so, I suspect that it will have to include in the analysis the interdependence of economic experience and political developments. It will have to treat at least some political phenomena not as independent variables—as exogenous variables in econometric jargon—but as themselves determined by economic events—as endogenous variables (Gordon 1975*b*). The second stage was greatly influenced by two major developments in economic theory of the past few decades—one, the analysis of imperfect information and of the cost of acquiring information, pioneered by George Stigler; the other, the role of human capital in determining the form of labor contracts, pioneered

by Gary Becker. The third stage will, I believe, be greatly influenced by a third major development—the application of economic analysis to political behavior, a field in which pioneering work has also been done by Stigler and Becker as well as by Kenneth Arrow, Duncan Black, Anthony Downs, James Buchanan, Gordon Tullock, and others.

The apparent positive relation between inflation and unemployment has been a source of great concern to government policymakers. Let me quote from a recent speech by Prime Minister Callaghan of Great Britain: “We used to think that you could just spend your way out of a recession and increase employment by cutting taxes and boosting Government spending. I tell you, in all candour, that that option no longer exists, and that insofar as it ever did exist, it only worked by injecting a bigger dose of inflation into the economy followed by higher levels of unemployment as the next step. That is the history of the past 20 years” (speech to Labour Party Conference, September 28, 1976).

The same view is expressed in a Canadian government white paper: “Continuing inflation, particularly in North America, has been accompanied by an increase in measured unemployment rates” (“The Way Ahead: A Framework for Discussion,” Government of Canada Working Paper, October 1976).

These are remarkable statements, running as they do directly counter to the policies adopted by almost every Western government throughout the postwar period.

A. *Some Evidence*

More systematic evidence for the past two decades is given in table 1 and figures 3 and 4, which show the rates of inflation and unemployment in seven industrialized countries over the past two decades. According to the 5-year averages in table 1, the rate of inflation and the level of unemployment moved in opposite directions—the expected simple Phillips curve outcome—in five out of seven countries between the first two quinquennia (1956–60, 1961–65); in only four out of seven countries between the second and third quinquennia (1961–65 and 1966–70); and in only one out of seven countries between the final two quinquennia (1966–70 and 1970–75). And even the one exception—Italy—is not a real exception. True, unemployment averaged a shade lower from 1971 to 1975 than in the prior 5 years, despite a more than tripling of the rate of inflation. However, since 1973 both inflation and unemployment have risen sharply.

The averages for all seven countries plotted in figure 3 bring out even more clearly the shift from a negatively sloped simple Phillips curve to a positively sloped one. The two curves move in opposite directions between the first two quinquennia; in the same direction thereafter.

TABLE 1
INFLATION AND UNEMPLOYMENT IN SEVEN COUNTRIES, 1956-75: AVERAGE VALUES FOR SUCCESSIVE QUINQUENNIA
(DP = RATE OF PRICE CHANGE, PERCENT PER YEAR; U = UNEMPLOYMENT, PERCENTAGE OF LABOR FORCE)

YEARS	FRANCE		GERMANY		ITALY		JAPAN		SWEDEN		UNITED KINGDOM		UNITED STATES		UNWEIGHTED AVERAGE SEVEN COUNTRIES	
	DP		DP		DP		DP		DP		DP		DP		DP	
	U		U		U		U		U		U		U		U	
1956 through 1960	5.6	1.1	1.8	2.9	1.9	6.7	1.9	1.4	3.7	1.9	2.6	1.5	2.0	5.2	2.8	3.0
1961 through 1965	3.7	1.2	2.8	0.7	4.9	3.1	6.2	0.9	3.6	1.2	3.5	1.6	1.3	5.5	3.7	2.0
1966 through 1970	4.4	1.7	2.4	1.2	3.0	3.5	5.4	1.1	4.6	1.6	4.6	2.1	4.2	3.9	4.1	2.2
1971 through 1975	8.8	2.5	6.1	2.1	11.3	3.3	11.4	1.4	7.9	1.8	13.0	3.2	6.7	6.1	9.3	2.9

NOTE.— DP is rate of change of consumer prices compounded annually from calendar year 1955 to 1960; 1960 to 1965; 1965 to 1970; 1970 to 1975; U is average unemployment during 5 indicated calendar years. As a result, DP is dated $\frac{1}{2}$ year prior to associated U .

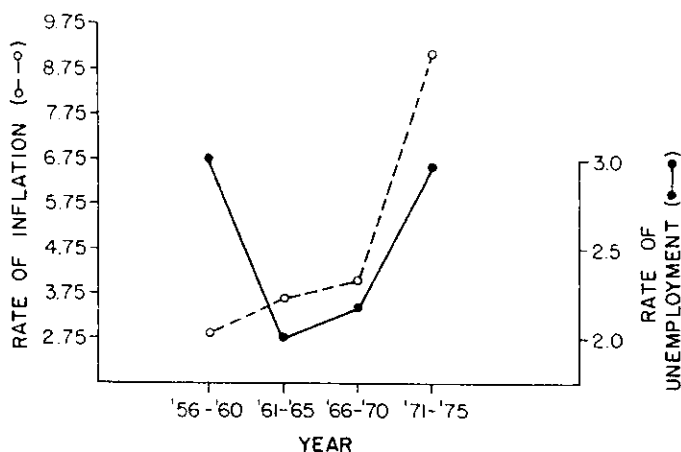


FIG. 3.—Rates of inflation and unemployment, 1956–75, by quinquennia; unweighted average for seven countries.

The annual data in figure 4 tell a similar, though more confused, story. In the early years, there is wide variation in the relation between prices and unemployment, varying from essentially no relation, as in Italy, to a fairly clear-cut year-to-year negative relation, as in the United Kingdom and the United States. In recent years, however, France, the United States, the United Kingdom, Germany, and Japan all show a clearly marked rise in both inflation and unemployment—though for Japan the rise in unemployment is much smaller relative to the rise in inflation than in the other countries, reflecting the different meaning of unemployment in the different institutional environment of Japan. Only Sweden and Italy fail to conform to the general pattern.

Of course, these data are at most suggestive. We do not really have seven independent bodies of data. Common international influences affect all countries so that multiplying the number of countries does not multiply proportionately the amount of evidence. In particular, the oil crisis hit all seven countries at the same time. Whatever effect the crisis had on the rate of inflation, it directly disrupted the productive process and tended to increase unemployment. Any such increases can hardly be attributed to the acceleration of inflation that accompanied them; at most the two could be regarded as at least partly the common result of a third influence (Gordon 1975*a*).

Both the quinquennial and annual data show that the oil crisis cannot wholly explain the phenomenon described so graphically by Callaghan. Already before the quadrupling of oil prices in 1973, most countries show a clearly marked association of rising inflation and rising unemployment. But this too may reflect independent forces rather than the

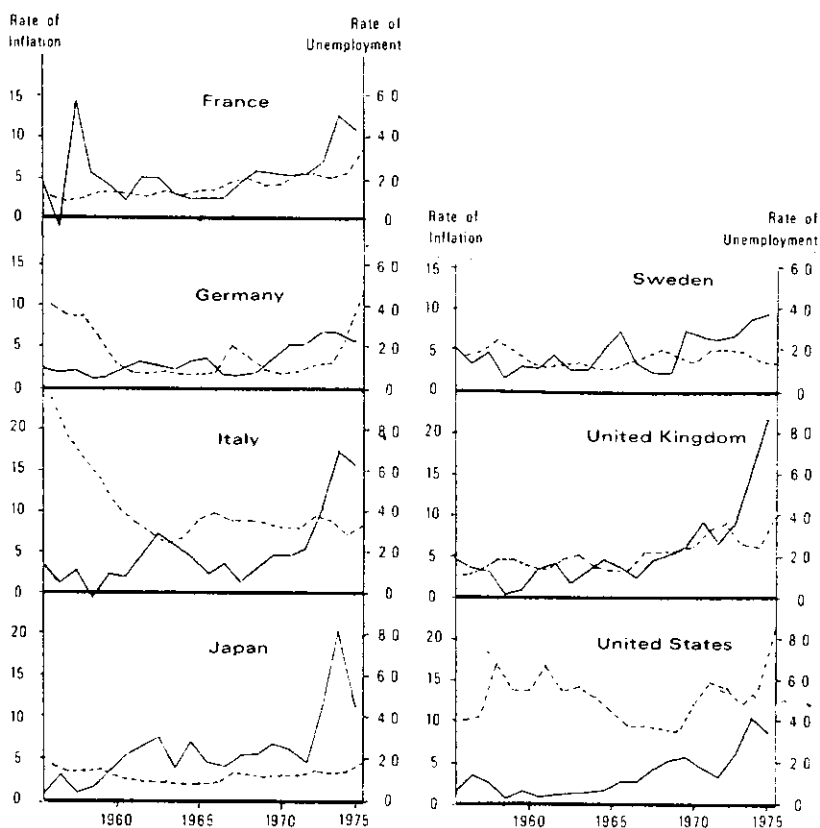


FIG. 4.—Inflation and unemployment in seven countries, annually, 1956–75. Solid line = rate of inflation; dashed line = rate of unemployment.

influence of inflation on unemployment. For example, the same forces that have been raising the natural rate of unemployment in the United States may have been operating in other countries and may account for their rising trend of unemployment, independently of the consequences of inflation.

Despite these qualifications, the data strongly suggest that, at least in some countries, of which Britain, Canada, and Italy may be the best examples, rising inflation and rising unemployment have been mutually reinforcing, rather than the separate effects of separate causes. The data are not inconsistent with the stronger statement that, in all industrialized countries, higher rates of inflation have some effects that, at least for a time, make for higher unemployment. The rest of this paper is devoted to a preliminary exploration of what some of these effects may be.

B. A Tentative Hypothesis

I conjecture that a modest elaboration of the natural-rate hypothesis is all that is required to account for a positive relation between inflation and unemployment, though of course such a positive relation may also occur for other reasons. Just as the natural-rate hypothesis explains a negatively sloped Phillips curve over short periods as a temporary phenomenon that will disappear as economic agents adjust their expectations to reality, so a positively sloped Phillips curve over somewhat longer periods may occur as a transitional phenomenon that will disappear as economic agents adjust not only their expectations but their institutional and political arrangements to a new reality. When this is achieved, I believe that—as the natural-rate hypothesis suggests—the rate of unemployment will be largely independent of the average rate of inflation, though the efficiency of utilization of resources may not be. High inflation need not mean either abnormally high or abnormally low unemployment. However, the institutional and political arrangements that accompany it, either as relics of earlier history or as products of the inflation itself, are likely to prove antithetical to the most productive use of employed resources—a special case of the distinction between the state of employment and the productivity of an economy referred to earlier.

Experience in many Latin American countries that have adjusted to chronically high inflation rates—experience that has been analyzed most perceptively by some of my colleagues, particularly Arnold Harberger (1967) and Larry Sjaastad (1974)—is consistent, I believe, with this view.

In the version of the natural-rate hypothesis summarized in figure 2, the vertical curve is for alternative rates of fully anticipated inflation. Whatever that rate—be it negative, zero or positive—it can be built into every decision if it is fully anticipated. At an anticipated 20 percent per year inflation, for example, long-term wage contracts would provide for a wage in each year that would rise relative to the zero-inflation wage by just 20 percent per year; long-term loans would bear an interest rate 20 percent higher than the zero-inflation rate or a principal that would be raised by 20 percent a year; and so on—in short, the equivalent of a full indexing of all contracts. The high rate of inflation would have some real effects, by altering desired cash balances, for example, but it need not alter the efficiency of labor markets, or the length or terms of labor contracts, and hence it need not change the natural rate of unemployment.

This analysis implicitly supposes, first, that inflation is steady or at least no more variable at a high rate than at a low—otherwise, it is unlikely that inflation would be as fully anticipated at high as at low

rates of inflation; second, that the inflation is, or can be, open, with all prices free to adjust to the higher rate, so that relative price adjustments are the same with a 20 percent inflation as with a zero inflation; third, really a variant of the second point, that there are no obstacles to indexing of contracts.

Ultimately, if inflation at an average rate of 20 percent per year were to prevail for many decades, these requirements could come fairly close to being met, which is why I am inclined to retain the long-long-run vertical Phillips curve. But when a country initially moves to higher rates of inflation, these requirements will be systematically departed from. And such a transitional period may well extend over decades.

Consider, in particular, the United States and the United Kingdom. For 2 centuries before World War II for the United Kingdom, and a century and a half for the United States, prices varied about a roughly constant level, showing substantial increases in time of war, then postwar declines to roughly prewar levels. The concept of a "normal" price level was deeply imbedded in the financial and other institutions of the two countries and in the habits and attitudes of their citizens.

In the immediate post-World War II period, prior experience was widely expected to recur. The fact was postwar inflation superimposed on wartime inflation; yet the expectation in both countries was deflation. It took a long time for the fear of postwar deflation to dissipate—if it still has—and still longer before expectations started to adjust to the fundamental change in the monetary system. That adjustment is still far from complete (Klein 1975).

Indeed, we do not know what a complete adjustment will consist of. We cannot know now whether the industrialized countries will return to the pre-World War II pattern of a long-term stable price level, or will move toward the Latin American pattern of chronically high inflation rates—with every now and then an acute outbreak of super- or hyper-inflation, as occurred recently in Chile and Argentina (Harberger 1976)—or will undergo more radical economic and political change leading to a still different resolution of the present ambiguous situation.

This uncertainty—or more precisely, the circumstances producing this uncertainty—leads to systematic departures from the conditions required for a vertical Phillips curve.

The most fundamental departure is that a high inflation rate is not likely to be steady during the transition decades. Rather, the higher the rate, the more variable it is likely to be. That has been empirically true of differences among countries in the past several decades (Jaffe and Kleiman 1975; Logue and Willett 1976). It is also highly plausible on theoretical grounds—both about actual inflation and, even more clearly, the anticipations of economic agents with respect to inflation. Governments have not produced high inflation as a deliberate announced

policy but as a consequence of other policies—in particular, policies of full employment and welfare-state policies raising government spending. They all proclaim their adherence to the goal of stable prices. They do so in response to their constituents, who may welcome many of the side effects of inflation but are still wedded to the concept of stable money. A burst of inflation produces strong pressure to counter it. Policy goes from one direction to the other, encouraging wide variation in the actual and anticipated rate of inflation. And, of course, in such an environment, no one has single-valued anticipations. Everyone recognizes that there is great uncertainty about what actual inflation will turn out to be over any specific future interval (Jaffe and Kleiman 1975; Meiselman 1976).

The tendency for inflation that is high on the average to be highly variable is reinforced by the effect of inflation on the political cohesiveness of a country in which institutional arrangements and financial contracts have been adjusted to a long-term “normal” price level. Some groups gain (e.g., homeowners); others lose (e.g., owners of savings accounts and fixed-interest securities). “Prudent” behavior becomes in fact reckless, and “reckless” behavior in fact prudent. The society is polarized; one group is set against another. Political unrest increases. The capacity of any government to govern is reduced at the same time that the pressure for strong action grows.

An increased variability of actual or anticipated inflation may raise the natural rate of unemployment in two rather different ways.

First, increased volatility shortens the optimum length of unindexed commitments and renders indexing more advantageous (Gray 1976). But it takes time for actual practice to adjust. In the meantime, prior arrangements introduce rigidities that reduce the effectiveness of markets. An additional element of uncertainty is, as it were, added to every market arrangement. In addition, indexing is, even at best, an imperfect substitute for stability of the inflation rate. Price indexes are imperfect; they are available only with a lag and generally are applied to contract terms only with a further lag.

These developments clearly lower economic efficiency. It is less clear what their effect is on recorded unemployment. High average inventories of all kinds are one way to meet increased rigidity and uncertainty. But that may mean labor hoarding by enterprises and low unemployment or a larger force of workers between jobs and so high unemployment. Shorter commitments may mean more rapid adjustment of employment to changed conditions and so low unemployment, or the delay in adjusting the length of commitments may lead to less satisfactory adjustment and so high unemployment. Clearly, much additional research is necessary in this area to clarify the relative importance of the various effects. About all one can say now is that the slow adjustment

of commitments and the imperfections of indexing may contribute to the recorded increase in unemployment.

A second related effect of increased volatility of inflation is to render market prices a less efficient system for coordinating economic activity. A fundamental function of a price system, as Hayek (1945) emphasized so brilliantly, is to transmit compactly, efficiently, and at low cost the information that economic agents need in order to decide what to produce and how to produce it, or how to employ owned resources. The relevant information is about *relative* prices—of one product relative to another, of the services of one factor of production relative to another, of products relative to factor services, of prices now relative to prices in the future. But the information in practice is transmitted in the form of *absolute* prices—prices in dollars or pounds or kronor. If the price level is on the average stable or changing at a steady rate, it is relatively easy to extract the signal about relative prices from the observed absolute prices. The more volatile the rate of general inflation, the harder it becomes to extract the signal about relative prices from the absolute prices: the broadcast about relative prices is, as it were, being jammed by the noise coming from the inflation broadcast (Lucas 1973, 1975; Harberger 1976). At the extreme, the system of absolute prices becomes nearly useless, and economic agents resort either to an alternative currency or to barter, with disastrous effects on productivity.

Again, the effect on economic efficiency is clear, on unemployment less so. But, again, it seems plausible that the average level of unemployment would be raised by the increased amount of noise in market signals, at least during the period when institutional arrangements are not yet adapted to the new situation.

These effects of increased volatility of inflation would occur even if prices were legally free to adjust—if, in that sense, the inflation were open. In practice, the distorting effects of uncertainty, rigidity of voluntary long-term contracts, and the contamination of price signals will almost certainly be reinforced by legal restrictions on price change. In the modern world, governments are themselves producers of services sold on the market: from postal services to a wide range of other items. Other prices are regulated by government and require government approval for change: from air fares to taxicab fares to charges for electricity. In these cases, governments cannot avoid being involved in the price-fixing process. In addition, the social and political forces unleashed by volatile inflation rates will lead governments to try to repress inflation in still other areas: by explicit price and wage control, or by pressuring private businesses or unions “voluntarily” to exercise “restraint,” or by speculating in foreign exchange in order to alter the exchange rate.

The details will vary from time to time and from country to country, but the general result is the same: reduction in the capacity of the price

system to guide economic activity; distortions in relative prices because of the introduction of greater friction, as it were, in all markets; and, very likely, a higher recorded rate of unemployment (Friedman 1976, chap. 12).

The forces I have just described may render the political and economic system dynamically unstable and produce hyperinflation and radical political change—as in many defeated countries after World War I, or in Chile and Argentina more recently. At the other extreme, before any such catastrophe occurs, policies may be adopted that will achieve a relatively low and stable rate of inflation and lead to the dismantling of many of the interferences with the price system. That would re-establish the preconditions for the straight-forward natural-rate hypothesis and enable that hypothesis to be used to predict the course of the transition.

An intermediate possibility is that the system will reach stability at a fairly constant though high average rate of inflation. In that case, unemployment should also settle down to a fairly constant level decidedly lower than during the transition. As the preceding discussion emphasizes, *increasing* volatility and *increasing* government intervention with the price system are the major factors that seem likely to raise unemployment, not *high* volatility or a *high* level of intervention.

Ways of coping with both volatility and intervention will develop: through indexing and similar arrangements for coping with volatility of inflation; through the development of indirect ways of altering prices and wages for avoiding government controls.

Under these circumstances, the long-run Phillips curve would again be vertical and we would be back at the natural-rate hypothesis, though perhaps for a different range of inflation rates than that for which it was first suggested.

Because the phenomenon to be explained is the coexistence of high inflation and high unemployment, I have stressed the effect of institutional changes produced by a transition from a monetary system in which there was a “normal” price level to a monetary system consistent with long periods of high, and possibly highly variable, inflation. It should be noted that once these institutional changes were made, and economic agents had adjusted their practices and anticipations to them, a reversal to the earlier monetary framework or even the adoption in the new monetary framework of a successful policy of low inflation would in its turn require new adjustments, and these might have many of the same adverse transitional effects on the level of employment. There would appear to be an intermediate-run negatively sloped Phillips curve instead of the positively sloped one I have tried to rationalize.

V. Conclusion

One consequence of the Keynesian revolution of the 1930s was the acceptance of a rigid absolute wage level, and a nearly rigid absolute price level, as a starting point for analyzing short-term economic change. It came to be taken for granted that these were essentially institutional data and were so regarded by economic agents, so that changes in aggregate nominal demand would be reflected almost entirely in output and hardly at all in prices. The age-old confusion between absolute prices and relative prices gained a new lease on life.

In this intellectual atmosphere it was understandable that economists would analyze the relation between unemployment and *nominal* rather than *real* wages and would implicitly regard changes in anticipated *nominal* wages as equal to changes in anticipated *real* wages. Moreover, the empirical evidence that initially suggested a stable relation between the level of unemployment and the rate of change of nominal wages was drawn from a period when, despite sharp short-period fluctuations in prices, there was a relatively stable long-run price level and when the expectation of continued stability was widely shared. Hence these data flashed no warning signals about the special character of the assumptions.

The hypothesis that there is a stable relation between the level of unemployment and the rate of inflation was adopted by the economics profession with alacrity. It filled a gap in Keynes's theoretical structure. It seemed to be the "one equation" that Keynes himself had said "we are . . . short" (1936, p. 276). In addition, it seemed to provide a reliable tool for economic policy, enabling the economist to inform the policy-maker about the alternatives available to him.

As in any science, so long as experience seemed to be consistent with the reigning hypothesis it continued to be accepted, although, as always, a few dissenters questioned its validity.

But as the '50s turned into the '60s, and the '60s into the '70s, it became increasingly difficult to accept the hypothesis in its simple form. It seemed to take larger and larger doses of inflation to keep down the level of unemployment. Stagflation reared its ugly head.

Many attempts were made to patch up the hypothesis by allowing for special factors such as the strength of trade unions. But experience stubbornly refused to conform to the patched-up versions.

A more radical revision was required. It took the form of stressing the importance of surprises—of differences between actual and anticipated magnitudes. It restored the primacy of the distinction between real and nominal magnitudes. There is a natural rate of unemployment at any time determined by real factors. This natural rate will tend to be

attained when expectations are on the average realized. The same real situation is consistent with any absolute level of prices or of price change, provided allowance is made for the effect of price change on the real cost of holding money balances. In this respect, money is neutral. On the other hand, unanticipated changes in aggregate nominal demand and in inflation will cause systematic errors of perception on the part of employers and employees alike that will initially lead unemployment to deviate in the opposite direction from its natural rate. In this respect, money is not neutral. However, such deviations are transitory, though it may take a long chronological time before they are reversed and finally eliminated as anticipations adjust.

The natural-rate hypothesis contains the original Phillips curve hypothesis as a special case and rationalizes a far broader range of experience, in particular the phenomenon of stagflation. It has by now been widely though not universally accepted.

However, the natural-rate hypothesis in its present form has not proved rich enough to explain a more recent development—a move from stagflation to slumpflation. In recent years, higher inflation has often been accompanied by higher unemployment—not lower unemployment, as the simple Phillips curve would suggest, nor the same unemployment, as the natural-rate hypothesis would suggest.

This recent association of higher inflation with higher unemployment may reflect the common impact of such events as the oil crisis, or independent forces that have imparted a common upward trend to inflation and unemployment.

However, a major factor in some countries and a contributing factor in others may be that they are in a transitional period—this time to be measured by quinquennia or decades not years. The public has not adapted its attitudes or its institutions to a new monetary environment. Inflation tends not only to be higher but also increasingly volatile and to be accompanied by widening government intervention into the setting of prices. The growing volatility of inflation and the growing departure of relative prices from the values that market forces alone would set combine to render the economic system less efficient, to introduce frictions in all markets, and, very likely, to raise the recorded rate of unemployment.

On this analysis, the present situation cannot last. It will either degenerate into hyperinflation and radical change, or institutions will adjust to a situation of chronic inflation, or governments will adopt policies that will produce a low rate of inflation and less government intervention into the fixing of prices.

I have told a perfectly standard story of how scientific theories are revised. Yet it is a story that has far-reaching importance.

Government policy about inflation and unemployment has been at

the center of political controversy. Ideological war has raged over these matters. Yet the drastic change that has occurred in economic theory has not been a result of ideological warfare. It has not resulted from divergent political beliefs or aims. It has responded almost entirely to the force of events: brute experience proved far more potent than the strongest of political or ideological preferences.

The importance for humanity of a correct understanding of positive economic science is vividly brought out by a statement made nearly two hundred years ago by Pierre S. du Pont, a deputy from Nemours to the French National Assembly, speaking, appropriately enough, on a proposal to issue additional assignats—the fiat money of the French Revolution: “Gentlemen, it is a disagreeable custom to which one is too easily led by the harshness of the discussions, to assume evil intentions. It is necessary to be gracious as to intentions; one should believe them good, and apparently they are; but we do not have to be gracious at all to inconsistent logic or to absurd reasoning. Bad logicians have committed more involuntary crimes than bad men have done intentionally” (September 25, 1790).

References

- Friedman, Milton. “The Methodology of Positive Economics.” In *Essays in Positive Economics*. Chicago: Univ. Chicago Press, 1953.
- . “An Inflationary Recession.” *Newsweek* (October 17, 1966). (a)
- . “What Price Guideposts?” In *Guidelines: Informal Contracts and the Market Place*, edited by G. P. Shultz and R. Z. Aliber. Chicago: Univ. Chicago Press, 1966. (b)
- . *Inflation: Causes and Consequences*. Bombay: Asia Publishing House, 1963. Reprinted in *Dollars and Deficits*. Englewood Cliffs, N.J.: Prentice-Hall, 1968. (a)
- . “The Role of Monetary Policy.” *A.E.R.* 58 (March 1968): 1–17. (b)
- . *Price Theory*. Chicago: Aldine, 1976.
- Gordon, Robert J. “Alternative Responses of Policy to External Supply Shocks.” *Brookings Papers Econ. Activity*, no. 1 (1975). (a)
- . “The Demand and Supply of Inflation.” *J. Law and Econ.* 18 (December 1975): 807–36. (b)
- . “Recent Developments in the Theory of Inflation and Unemployment.” *J. Monetary Econ.* 2 (1976): 185–219.
- Gray, Jo Anna. “Essays on Wage Indexation.” Ph.D. dissertation, Univ. Chicago, 1976.
- Harberger, Arnold G. “The Inflation Problem in Latin America.” A report prepared for the Buenos Aires (March 1966) meeting of the Inter-American Committee of the Alliance for Progress, published in Spanish as “El problema de la inflación en America Latina,” in Centro de Estudios Monetarios Latinoamericanos, *Boletín mensual* (June 1966), pp. 253–69. Reprinted in Economic Development Institute, *Trabajos sobre desarrollo económico*. Washington: IBRD, 1967.

- . "Inflation." In *The Great Ideas Today, 1976*, edited by Robert M. Hutchins and Mortimer J. Adler. Chicago: Encyclopaedia Britannica, 1976.
- Hayek, F. A. "The Use of Knowledge in Society." *A.E.R.* 35 (September 1945): 519–30.
- Jaffe, Dwight, and Kleiman, Ephraim. "The Welfare Implications of Uneven Inflation." Seminar paper no. 50, Inst. Internat. Econ. Studies, Univ. Stockholm, November 1975.
- Keynes, J. M. *General Theory of Employment, Interest, and Money*. London: Macmillan, 1936.
- Klein, Benjamin. "Our New Monetary Standard: The Measurement and Effects of Price Uncertainty, 1880–1973." *Econ. Inquiry* 13 (December 1975): 461–84.
- Logue, Dennis E., and Willett, Thomas D. "A Note on the Relation between the Rate and Variability of Inflation." *Economica* 43 (May 1976): 151–58.
- Lucas, Robert E. "Some International Evidence on Output-Inflation Trade-offs." *A.E.R.* 63 (June 1973): 326–34.
- . "An Equilibrium Model of the Business Cycle." *J.P.E.* 83, no. 6 (December 1975): 1113–44.
- Meiselman, David. "Capital Formation, Monetary and Financial Adjustments." *Proceedings, 27th National Conference of Tax Foundation*. Washington: Tax Foundation, 1976.
- Muth, John. "Rational Expectations and the Theory of Price Movements." *Econometrica* 29 (July 1961): 315–33.
- Phelps, E. S. "Phillips Curve, Expectations of Inflation and Optimal Unemployment over Time." *Economica*, n.s. 34 (August 1967): 254–81.
- . "Money Wage Dynamics and Labor Market Equilibrium." In *Microeconomic Foundations of Employment and Inflation Theory*, edited by E. S. Phelps. New York: Norton, 1970.
- Phillips, A. W. "The Relationship between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957." *Economica* (November 1958), pp. 283–99.
- Sjaastad, Larry A. "Monetary Policy and Suppressed Inflation in Latin America." In *National Monetary Policies and the International Financial System*, edited by Robert Z. Aliber. Chicago: Univ. Chicago Press, 1974.

The Utility Analysis of Choices Involving Risk¹

with L. J. Savage

I. The Problem and Its Background

The purpose of this paper is to suggest that an important class of reactions of individuals to risk can be rationalized by a rather simple extension of orthodox utility analysis.

Individuals frequently must, or can, choose among alternatives that differ, among other things, in the degree of risk to which the individual will be subject. The clearest examples are provided by insurance and gambling. An individual who buys fire insurance on a house he owns is accepting the certain loss of a small sum (the insurance premium) in preference to the combination of a small chance of a much larger loss (the value of the house) and a large chance of no loss. That is, he is choosing certainty in preference to uncertainty. An individual who buys a lottery ticket is subjecting himself to a large chance of losing a small amount (the price of the lottery ticket) plus a small chance of winning a large amount (a prize) in preference to avoiding both risks. He is choosing uncertainty in preference to certainty.

This choice among different degrees of risk so prominent in insurance and gambling, is clearly present and important in a much broader range of economic choices. Occupations differ greatly in the variability of the income they promise: in some, for example, civil service employment, the prospective income is rather clearly defined and is almost certain to be within rather narrow limits; in others, for example, salaried employment as an accountant, there is somewhat more variability yet almost no chance of either an extremely high or an extremely low income; in still others, for example, motion-picture acting, there is extreme variability, with a small chance of an extremely high income and a larger

¹ The fundamental ideas of this paper were worked out jointly by the two authors. The paper was written primarily by the senior author.

chance of an extremely low income. Securities vary similarly, from government bonds and industrial "blue chips" to "blue-sky" common stocks; and so do business enterprises or lines of business activity. Whether or not they realize it and whether or not they take explicit account of the varying degree of risk involved, individuals choosing among occupations, securities, or lines of business activity are making choices analogous to those that they make when they decide whether to buy insurance or to gamble. Is there any consistency among the choices of this kind that individuals make? Do they neglect the element of risk? Or does it play a central role? If so, what is that role?

These problems have, of course, been considered by economic theorists, particularly in their discussions of earnings in different occupations and of profits in different lines of business.² Their treatment of these problems has, however, never been integrated with their explanation of choices among riskless alternatives. Choices among riskless alternatives are explained in terms of maximization of utility: individuals are supposed to choose as they would if they attributed some common quantitative characteristic—designated utility—to various goods and then selected the combination of goods that yielded the largest total amount of this common characteristic. Choices among alternatives involving different degrees of risk, for example, among different occupations, are explained in utterly different terms—by ignorance of the odds or by the fact that "young men of an adventurous disposition are more attracted by the prospects of a great success than they are deterred by the fear of failure," by "the overweening conceit which the greater part of men have of their own abilities," by "their absurd presumption in their own good fortune," or by some similar *deus ex machina*.³

The rejection of utility maximization as an explanation of choices among different degrees of risk was a direct consequence of the belief in diminishing marginal utility. If the marginal utility of money diminishes, an individual seeking to maximize utility will never participate in a "fair" game of chance, for example, a game in which he has an equal chance of winning or losing a dollar. The gain in utility from winning a dollar will be less than the loss in utility from losing a dollar, so that the expected utility from participation in the game is negative. Diminishing marginal utility plus maximization of expected utility would thus imply that individuals would always have to be paid to induce them to

² E.g., see Adam Smith, *The Wealth of Nations*, Book 1, chap. x (Modern Library reprint of Cannan ed.), pp. 106–11; Alfred Marshall, *Principles of Economics* (8th ed.; London: Macmillan & Co., Ltd., 1920), pp. 398–400, 554–55, 613.

³ Marshall, *op. cit.*, p. 554 (first quotation); Smith, *op. cit.*, p. 107 (last two quotations).

bear risk.⁴ But this implication is clearly contradicted by actual behavior. People not only engage in fair games of chance, they engage freely and often eagerly in such unfair games as lotteries. Not only do risky occupations and risky investments not always yield a higher average return than relatively safe occupations or investments, they frequently yield a much lower average return.

Marshall resolved this contradiction by rejecting utility maximization as an explanation of choices involving risk. He need not have done so, since he did not need diminishing marginal utility—or, indeed, any quantitative concept of utility—for the analysis of riskless choices. The shift from the kind of utility analysis employed by Marshall to the indifference-curve analysis of F. Y. Edgeworth, Irving Fisher, and Vilfredo Pareto revealed that to rationalize riskless choices, it is sufficient to suppose that individuals can rank baskets of goods by total utility. It is unnecessary to suppose that they can compare differences between utilities. But diminishing, or increasing, marginal utility implies a comparison of differences between utilities and hence is an entirely gratuitous assumption in interpreting riskless choices.

The idea that choices among alternatives involving risk can be explained by the maximization of expected utility is ancient, dating back at least to D. Bernoulli's celebrated analysis of the St. Petersburg paradox.⁵ It has been repeatedly referred to since then but almost invariably rejected as the correct explanation—commonly because the prevailing belief in diminishing marginal utility made it appear that the existence of gambling could not be so explained. Even since the widespread

⁴ See Marshall, *op. cit.*, p. 135 n.; Mathematical Appendix, n. ix (p. 843). "Gambling involves an economic loss, even when conducted on perfectly fair and even terms. . . . A theoretically fair insurance against risks is always an economic gain" (p. 135). "The argument that fair gambling is an economic blunder . . . requires no further assumption than that, firstly the pleasures of gambling may be neglected; and, secondly $\phi''(x)$ is negative for all values of x , where $\phi(x)$ is the pleasure derived from wealth equal to x It is true that this loss of probable happiness need not be greater than the pleasure derived from the excitement of gambling, and we are then thrown back upon the induction that pleasures of gambling are in Bentham's phrase 'impure'; since experience shows that they are likely to engender a restless, feverish character, unsuited for steady work as well as for the higher and more solid pleasures of life" (p. 843).

⁵ See Daniel Bernoulli, *Versuch einer neuen Theorie der Wertbestimmung von Glücksfällen* (Leipzig, 1896), translated by A. Pringsheim from "Specimen theoriae novae de mensura sortis," *Commentarii academiae scientiarum imperialis Petropolitanae*, Vol. V, for the years 1730 and 1731, published in 1738.

In an interesting note appended to his paper Bernoulli points out that Cramer (presumably Gabriel Cramer [1704–52]), a famous mathematician of the time, had anticipated some of his own views by a few years. The passages that he quotes from a letter in French by Cramer contain what, to us, is the truly essential point in Bernoulli's paper, namely, the idea of using the mathematical expectation of utility (the "moral expectation") instead of the mathematical expectation of income to compare alternatives involving risk. Cramer has not in general been attributed this much credit, apparently because the essential point in Bernoulli's paper has been taken to be the suggestion that the logarithm of income is an appropriate utility function.

recognition that the assumption of diminishing marginal utility is unnecessary to explain riskless choices, writers have continued to reject maximization of expected utility as "unrealistic."⁶ This rejection of maximization of expected utility has been challenged by John von Neumann and Oskar Morgenstern in their recent book, *Theory of Games and Economic Behavior*.⁷ They argue that "under the conditions on which the indifference curve analysis is based very little extra effort is needed to reach a numerical utility," the expected value of which is maximized in choosing among alternatives involving risk.⁸ The present paper is based on their treatment but has been made self-contained by the paraphrasing of essential parts of their argument.

If an individual shows by his market behavior that he prefers *A* to *B* and *B* to *C*, it is traditional to rationalize this behavior by supposing that he attaches more utility to *A* than to *B* and more utility to *B* than to *C*. All utility functions that give the same ranking to possible alternatives will provide equally good rationalizations of such choices, and it will make no difference which particular one is used. If, in addition, the individual should show by his market behavior that he prefers a 50-50 chance of *A* or *C* to the certainty of *B*, it seems natural to rationalize this behavior by supposing that the *difference* between the utilities he attaches to *A* and *B* is greater than the *difference* between the utilities he attaches to *B* and *C*, so that the *expected* utility of the preferred combination is greater than the utility of *B*. The class of utility functions, if there be any, that can provide the same ranking of alternatives that involve risk is much more restricted than the class that can provide the same ranking of alternatives that are certain. It consists of utility functions that differ only in origin and unit of measure (i.e., the utility

⁶ "It has been the assumption in the classical literature on this subject that the individual in question will always try to maximize the mathematical expectation of his gain or utility. . . . This may appear plausible, but it is certainly not an assumption which must hold true in all cases. It has been pointed out that the individual may also be interested in, and influenced by, the range or the standard deviation of the different possible utilities derived or some other measure of dispersion. It appears pretty evident from the behavior of people in lotteries or football pools that they are not a little influenced by the skewness of the probability distribution" (Gerhard Tintner, "A Contribution to the Non-Static Theory of Choice," *Quarterly Journal of Economics*, LVI [February, 1942], 278).

"It would be definitely unrealistic . . . to confine ourselves to the mathematical expectation only, which is the usual but not justifiable practice of the traditional calculus of 'moral probabilities'" (J. Marschak, "Money and the Theory of Assets," *Econometrica*, VI [1938], 320).

Tintner's inference, apparently also shared by Marschak, that the facts he cites are necessarily inconsistent with maximization of expected utility is erroneous (see secs. 3 and 4 below). He is led to consider a formally more general solution because of his failure to appreciate the real generality of the kinds of behavior explicable by the maximization of expected utility.

⁷ Princeton University Press, 1st ed., 1944; 2d ed., 1947; pp. 15-31 (both eds.), pp. 617-32 (2d ed. only); succeeding references are to 2d ed.

⁸ *Ibid.*, p. 17.

functions in the class are linear functions of one another).⁹ Thus, in effect, the ordinal properties of utility functions can be used to rationalize riskless choices, the numerical properties to rationalize choices involving risk.

It does not, of course, follow that there will exist a utility function that will rationalize in this way the reactions of individuals to risk. It may be that individuals behave inconsistently—sometimes choosing a 50-50 chance of *A* or *C* instead of *B* and sometimes the reverse; or sometimes choosing *A* instead of *B*, *B* instead of *C*, and *C* instead of *A*—or that in some other way their behavior is different from what it would be if they were seeking rationally to maximize expected utility in accordance with a given utility function. Or it may be that some types of reactions to risk can be rationalized in this way while others cannot. Whether a numerical utility function will in fact serve to rationalize any particular class of reactions to risk is an empirical question to be tested; there is no obvious contradiction such as was once thought to exist.

This paper attempts to provide a crude empirical test by bringing together a few broad observations about the behavior of individuals in choosing among alternatives involving risk (sec. 2) and investigating whether these observations are consistent with the hypothesis revived by von Neumann and Morgenstern (secs. 3 and 4). It turns out that these empirical observations are entirely consistent with the hypothesis if a rather special shape is given to the total utility curve of money (sec. 4). This special shape, which can be given a tolerably satisfactory interpretation (sec. 5), not only brings under the aegis of rational utility maximization much behavior that is ordinarily explained in other terms but also has implications about observable behavior not used in deriving it (sec. 6). Further empirical work should make it possible to determine whether or not these implications conform to reality.

It is a testimony to the strength of the belief in diminishing marginal utility that it has taken so long for the possibility of interpreting gambling and similar phenomena as a contradiction of universal diminishing marginal utility, rather than of utility maximization, to be recognized. The initial mistake must have been at least partly a product of a strong introspective belief in diminishing marginal utility: a dollar must mean less to a rich man than to a poor man; see how much more a man will spend when he is rich than when he is poor to avoid any given amount of pain or discomfort.¹⁰ Some of the comments that have been published

⁹ *Ibid.*, pp. 15–31, esp. p. 25.

¹⁰ This elemental argument seems so clearly to justify diminishing marginal utility that it may be desirable even now to state explicitly how this phenomenon can be rationalized equally well on the assumption of increasing marginal utility of money. It is only necessary to suppose that the avoidance of pain and the other goods that can be bought with money are related goods and that, while the marginal utility of money increases as the amount of money increases, the marginal utility of avoiding pain increases even faster.

by competent economists on the utility analysis of von Neumann and Morgenstern are even more remarkable testimony to the hold that diminishing marginal utility has on economists. Vickrey remarks: "There is abundant evidence that individual decisions in situations involving risk are not always made in ways that are compatible with the assumption that the decisions are made rationally with a view to maximizing the mathematical expectation of a utility function. The purchase of tickets in lotteries, sweepstakes, and 'numbers' pools would imply, on such a basis, that the marginal utility of money is an increasing rather than a decreasing function of income. Such a conclusion is obviously unacceptable as a guide to social policy."¹¹ Kaysen remarks, "Unfortunately, these postulates [underlying the von Neumann and Morgenstern discussion of utility measurement] involve an assumption about economic behavior which is contrary to experience. . . . That this assumption is contradicted by experience can easily be shown by hundreds of examples [including] the participation of individuals in lotteries in which their mathematical expectation of gain (utility) is negative."¹²

2. Observable Behavior to Be Rationalized

The economic phenomena to which the hypothesis revived by von Neumann and Morgenstern is relevant can be divided into, first, the phenomena ordinarily regarded as gambling and insurance; second, other economic phenomena involving risk. The latter are clearly the more important, and the ultimate significance of the hypothesis will depend primarily on the contribution it makes to an understanding of them. At the same time, the influence of risk is revealed most markedly in gambling and insurance, so that these phenomena have a significance for testing and elaborating the hypothesis out of proportion to their importance in actual economic behavior.

At the outset it should be confessed that we have conducted no extensive empirical investigation of either class of phenomena. For the present, we are content to use what is already available in the literature, or obvious from casual observation, to provide a first test of the hypothesis and to impose significant substantive restrictions on it.

The major economic decisions of an individual in which risk plays an important role concern the employment of the resources he controls:

¹¹ William Vickrey, "Measuring Marginal Utility by Reactions to Risk," *Econometrica*, XIII (1945), 319–33. The quotation is from pp. 327 and 328. "The purchase of tickets in lotteries, sweepstakes, and 'numbers' pools" does not imply that marginal utility of money increases with income everywhere (see sec. 4 below). Moreover, it is entirely unnecessary to identify the quantity that individuals are to be interpreted as maximizing with a quantity that should be given special importance in public policy.

¹² C. Kaysen, "A Revolution in Economic Theory?" *Review of Economic Studies*, XIV, No. 35 (1946–47), 1–15; quotation is from p. 13.

what occupation to follow, what entrepreneurial activity to engage in, how to invest (nonhuman) capital. Alternative possible uses of resources can be classified into three broad groups according to the degree of risk involved: (*a*) those involving little or no risk about the money return to be received—occupations like schoolteaching, other civil service employment, clerical work; business undertakings of a standard, predictable type like many public utilities; securities like government bonds, high-grade industrial bonds; some real property, particularly owner-occupied housing; (*b*) those involving a moderate degree of risk but unlikely to lead to either extreme gains or extreme losses—occupations like dentistry, accountancy, some kinds of managerial work; business undertakings of fairly standard kinds in which, however, there is sufficient competition to make the outcome fairly uncertain; securities like lower-grade bonds, preferred stocks, higher-grade common stocks; (*c*) those involving much risk, with some possibility of extremely large gains and some of extremely large losses—occupations involving physical risks, like piloting aircraft, automobile racing, or professions like medicine and law; business undertakings in untried fields; securities like highly speculative stocks; some types of real property.

The most significant generalization in the literature about choices among these three uses of resources is that, other things the same, uses *a* or *c* tend in general to be preferred to use *b*; that is, people must in general be paid a premium to induce them to undertake moderate risks instead of subjecting themselves to either small or large risks. Thus Marshall says: "There are many people of a sober steady-going temper, who like to know what is before them, and who would far rather have an appointment which offered a certain income of say £400 a year than one which was not unlikely to yield £600, but had an equal chance of affording only £200. Uncertainty, therefore, which does not appeal to great ambitions and lofty aspirations, has special attractions for very few; while it acts as a deterrent to many of those who are making their choice of a career. And as a rule the certainty of moderate success attracts more than an expectation of an uncertain success that has an equal actuarial value.

"But on the other hand, if an occupation offers a few extremely high prizes, its attractiveness is increased out of all proportion to their aggregate value."¹³

Adam Smith comments similarly about occupational choices and, in addition, says of entrepreneurial undertakings: "The ordinary rate of profits always rises more or less with the risk. It does not, however, seem to rise in proportion to it, or so as to compensate it completely. . . . The presumptuous hope of success seems to act here as upon all other

¹³ *Op. cit.*, pp. 554–55.

occasions, and to entice so many adventurers into those hazardous trades, that their competition reduces the profit below what is sufficient to compensate the risk.”¹⁴

Edwin Cannan, in discussing the rate of return on investments, concludes that “the probability is that the classes of investments which on the average return most to the investor are neither the very safest of all nor the very riskiest, but the intermediate classes which do not appeal either to timidity or to the gambling instinct.”¹⁵

This asserted preference for extremely safe or extremely risky investments over investments with an intermediate degree of risk has its direct counterpart in the willingness of persons to buy insurance and also to buy lottery tickets or engage in other forms of gambling involving a small chance of a large gain. The extensive market for highly speculative stocks—the kind of stocks that “blue-sky” laws are intended to control—is a border-line case that could equally well be designated as investment or gambling.

The empirical evidence for the willingness of persons of all income classes to buy insurance is extensive.¹⁶ Since insurance companies have

¹⁴ *Op. cit.*, p. 111.

¹⁵ Article on “Profit,” in *Dictionary of Political Economy*, ed. R. H. Inglis Palgrave (new edition, ed. Henry Higgs; London, 1926); see also the summary of the views of different writers on risk-taking in F. H. Knight, *Risk, Uncertainty, and Profit* (New York, 1921; reprint London School of Economics and Political Science, 1933), pp. 362–67.

¹⁶ E.g., see U.S. Bureau of Labor Statistics, *Bulletin 648: Family Expenditures in Selected Cities, 1935–36*, Vol. I: *Family Expenditures for Housing, 1935–36*; Vol. VI: *Family Expenditures for Transportation, 1935–36*; and Vol. VIII: *Changes in Assets and Liabilities, 1935–36*.

Table 6 of the Tabular Summary of Vol. I gives the percentage of home-owning families reporting the payment of premiums for insurance on the house. These percentages are given separately for each income class in each of a number of cities or groups of cities. Since premiums are often paid less frequently than once a year, the percentages given definitely understate the percentage of families carrying insurance. Yet the bulk of the percentages are well over 40.

Table 5 of the Tabular Summary of Vol. VI gives the percentage of families (again by income classes and cities or groups of cities) reporting expenditures for automobile insurance. These figures show a very rapid increase in the percentage of automobile operators that had insurance (this figure is derived by dividing the percentage of families reporting automobile insurance by the percentage of families operating cars) as income increases. In the bottom income classes, where operation of a car is infrequent, only a minority of those who operate cars carry insurance. In the upper income classes, where most families operate cars, the majority of operators carry insurance. A convenient summary of these percentages for selected income classes in six large cities, given in text Table 10 (p. 26), has forty-two entries. These vary from 4 per cent to 98 per cent and twenty-three are over 50 per cent.

Table 3 of the Tabular Summary of Vol. VIII gives the percentage of families in each income class in various cities or groups of cities reporting the payment of life, endowment, or annuity insurance premiums. The percentages are uniformly high. For example, for New York City the percentage of white families reporting the payment of insurance premiums is 75 per cent or higher for every income class listed and varies from 75 per cent in the income class \$500–\$749 to over 95 per cent in the upper-income classes; the percentage of Negro families purchasing insurance was 38 percent for the \$1,000–\$1,249

costs of operation that are covered by their premium receipts, the purchaser is obviously paying a larger premium than the average compensation he can expect to receive for the losses against which he carries insurance. That is, he is paying something to escape risk.

The empirical evidence for the willingness of individuals to purchase lottery tickets, or engage in similar forms of gambling, is also extensive. Many governments find, and more governments have found, lotteries an effective means of raising revenue.¹⁷ Though illegal, the "numbers"

class but 60 per cent or higher for every other class. This story is repeated for city after city, the bulk of the entries in the table for the percentage of families purchasing insurance being above 80 per cent.

These figures cannot be regarded as direct estimates of the percentage of families willing to pay something—that is, to accept a smaller actuarial value—in order to escape risk, the technical meaning of the purchase of insurance that is relevant for our purpose. (1) The purchase of automobile and housing insurance may not be a matter of choice. Most owned homes have mortgages (see I, 361, Table L) and the mortgage may require that insurance be carried. The relevant figure for mortgaged homes would be the fraction of owners carrying a larger amount of insurance than is required by the mortgage. Similarly, finance companies generally require that insurance be carried on automobiles purchased on the instalment plan and not fully paid for, and the purchase of automobile insurance is compulsory in some states. (2) For automobile property damage and liability insurance (but not collision insurance) the risks to the operator and to the insurance company may not be the same, particularly to persons in the lower-income classes. The loss to the uninsured operator is limited by his wealth and borrowing power, and the maximum amount that he can lose may be well below the face value of the policy that he would purchase. The excess of the premium over the expected loss is thus greater for him than for a person with more wealth or borrowing power. The rise in the percentage of persons carrying automobile insurance as income rises may therefore reflect not an increased willingness to carry insurance but a reduction in the effective price that must be paid for insurance. (3) This tendency may be reversed for the relatively high-income classes for both automobile and housing insurance by the operation of the income tax. Uninsured losses are in many instances deductible from income before computation of income tax under the United States federal income tax, while insurance premiums are not. This tends to make the net expected loss less for the individual than for the insurance company. This effect is almost certainly negligible for the figures cited above, both because they do not effectively cover very high incomes and because the federal income tax was relatively low in 1935–36. (4) Life insurance at times comes closer to gambling (the choice of an uncertain alternative in preference to a certain alternative with a higher expected value) than to the payment of a premium to escape risk. For example, special life-insurance policies purchased to cover a single railroad or airplane trip are probably more nearly comparable to a lottery ticket than a means of achieving certainty. (5) Even aside from these qualifications, actual purchase of insurance would give at best a lower limit to the number willing to buy insurance, since there will always be some who will regard the price asked as too high.

These qualifications offset one another to some extent. It seems highly unlikely that their net effect could be sufficient to reverse the conclusion suggested by the evidence cited that a large fraction of people in all income classes are willing to buy insurance.

¹⁷ France, Spain, and Mexico, to name but three examples, currently conduct lotteries for revenue. Russia attaches a lottery feature to bonds sold to the public. Great Britain conducted lotteries from 1694 to 1826. In the United States lotteries were used extensively before the Revolution and for some time thereafter, both directly by state governments and under state charters granted to further specific projects deemed to have a state interest. For the history of lotteries in Great Britain see C. L'Estrange Ewen, *Lotteries and Sweepstakes*

game and similar forms of gambling are reported to flourish in the United States,¹⁸ particularly among the lower income classes.

It seems highly unlikely that there is a sharp dichotomy between the individuals who purchase insurance and those who gamble. It seems much more likely that many do both or, at any rate, would be willing to. We can cite no direct evidence for this asserted fact, though indirect evidence and casual observation give us considerable confidence that it is correct. Its validity is suggested by the extensiveness of both gambling and the purchase of insurance. It is also suggested by some of the available evidence on how people invest their funds. The widespread legislation against "bucket shops" suggests that relatively poor people must have been willing to buy extremely speculative stocks of a "blue-sky" variety. Yet the bulk of the property income of the lower-income classes consists of interest and rents and relatively little of dividends, whereas the reverse is true for the upper-income classes.¹⁹ Rents and interest are types of receipts that tend to be derived from investments with relatively little risk, and so correspond to the purchase of insurance, whereas investment in speculative stocks corresponds to the purchase of lottery tickets.

Offhand it appears inconsistent for the same person both to buy insurance and to gamble: he is willing to pay a premium, in the one case, to avoid risk, in the other, to bear risk. And indeed it would be inconsistent for a person to be willing to pay something (no matter how little) in excess of actuarial value to avoid every possible risk and also something in excess of actuarial value to assume every possible risk. One must distinguish among different kinds of insurance and different kinds of gambling, since a willingness to pay something for only some kinds of insurance would not necessarily be inconsistent with a willingness to engage in only some kinds of gambling. Unfortunately, very little empirical evidence is readily available on the kinds of insurance that people are willing to buy and the kinds of gambling that they are willing to

(London, 1932); in New York State, A. F. Ross, "History of Lotteries in New York," *Magazine of History*, Vol. V (New York, 1907). There seem to be no direct estimates of the fraction of the people who purchase tickets in state or other legal lotteries, and it is clear that such figures would be difficult to get from data obtained in connection with running the lotteries. The receipts from legal lotteries, and casual impressions of observers, suggest that a substantial fraction of the relevant units (families or, alternatively, individual income recipients) purchase tickets.

¹⁸ Evidence from wagering on horse races, where this has been legalized, is too ambiguous to be of much value. Since most legal wagering is at the track, gambling is available only to those who go to watch the races and is combined with participation in the mechanics of the game of chance.

¹⁹ *Delaware Income Statistics*, Vol. I (Bureau of Economic and Business Research, University of Delaware, 1941), Table 1; *Minnesota Incomes, 1938-39*, Vol. II (Minnesota Resources Commission, 1942), Table 27; F. A. Hanna, J. A. Pechman, S. M. Lerner, *Analysis of Wisconsin Income* ("Studies in Income and Wealth," Vol. IX [National Bureau of Economic Research, 1948]), Part II, Table 1.

engage in. About the only clear indication is that people are willing to enter into gambles that offer a small chance of a large gain—as in lotteries and “blue-sky” securities.

Lotteries seem to be an extremely fruitful, and much neglected, source of information about reactions of individuals to risk. They present risk in relatively pure form, with little admixture of other factors; they have been conducted in many countries and for many centuries, so that a great deal of evidence is available about them; there has been extensive experimentation with the terms and conditions that would make them attractive, and much competition in conducting them, so that any regularities they may show would have to be interpreted as reflecting corresponding regularities in human behavior.²⁰ It is, of course, not certain that inferences from lotteries would carry over to other choices involving risk. There would, however, seem to be some presumption that they would do so, though of course the validity of this presumption would have to be tested.²¹

The one general feature of lotteries that is worth noting in this preliminary survey, in addition to the general willingness of people to participate in them, is the structure of prizes that seems to have developed. Lotteries rarely have just a single prize equal to the total sum to be paid out as prizes. Instead, they tend to have several or many prizes. The largest prize is ordinarily not very much larger than the next largest, and often there is not one largest prize but several of the same size.²² This tendency is so general that one would expect it to reflect some consistent feature of individual reactions, and any hypothesis designed to explain reactions to uncertainty should explain it.

3. The Formal Hypothesis

The hypothesis that is proposed for rationalizing the behavior just summarized can be stated compactly as follows: In choosing among alternatives open to it, whether or not these alternatives involve risk, a consumer unit (generally a family, sometimes an individual) behaves as if (a) it had a consistent set of preferences; (b) these preferences could be completely described by a function attaching a numerical value—to

²⁰ Aside from their value in providing information about reactions to risk, data from lotteries may be of broader interest in providing evidence about the stability of tastes and preferences over time and their similarity in different parts of the world. Here is a “commodity” which has remained unchanged over centuries, which is the same all over the globe, and which has been dealt in widely for the entire period and over much of the globe. It is hard to conceive of any other commodity for which this is true.

²¹ See Smith, *op. cit.*, p. 108, for a precedent.

²² See Ewen, *op. cit.*, *passim*, but esp. descriptions of state lotteries in chap. vii, pp. 199–244; see also the large numbers of bills advertising lotteries in John Ashton, *A History of English Lotteries* (London: Leadenhall Press, 1893).

be designated "utility"—to alternatives each of which is regarded as certain; (c) its objective were to make its expected utility as large as possible. It is the contribution of von Neumann and Morgenstern to have shown that an alternative statement of the same hypothesis is: An individual chooses in accordance with a system of preferences which has the following properties:

1. The system is complete and consistent; that is, an individual can tell which of two objects he prefers or whether he is indifferent between them, and if he does not prefer *C* to *B* and does not prefer *B* to *A*, then he does not prefer *C* to *A*.²³ (In this context, the word "object" includes combinations of objects with stated probabilities; for example, if *A* and *B* are objects, a 40-60 chance of *A* or *B* is also an object.)
2. Any object which is a combination of other objects with stated probabilities is never preferred to every one of these other objects, nor is every one of them ever preferred to the combination.
3. If the object *A* is preferred to the object *B* and *B* to the object *C*, there will be some probability combination of *A* and *C* such that the individual is indifferent between it and *B*.²⁴

This form of statement is designed to show that there is little difference between the plausibility of this hypothesis and the usual indifference-curve explanation of riskless choices.

These statements of the hypothesis conceal by their very compactness most of its implications. It will pay us, therefore, to elaborate them. It simplifies matters, and involves no loss in generality, to regard the alternatives open to the consumer unit as capable of being expressed entirely in terms of money or money income. Actual alternatives are not, of course, capable of being so expressed: the same money income may be valued very differently according to the terms under which it is to be received, the nonpecuniary advantages or disadvantages associated

²³ The transitivity of the relation of indifference assumed in this postulate is, of course, an idealization. It is clearly possible that the difference between successive pairs of alternatives in a series might be imperceptible to an individual, yet the first of the series definitely preferable to the last. This idealization, which is but a special case of the idealization involved in the geometric concept of a dimensionless point, seems to us unobjectionable. However, the use of this idealization in indifference-curve analysis is the major criticism offered by W. E. Armstrong in an attack on indifference-curve analysis in his article "The Determinateness of the Utility Function," *Economic Journal*, XLIX (September, 1939), 453–67. In a more recent article ("Uncertainty and the Utility Function," *Economic Journal*, LVIII [March, 1948], 1–10) Armstrong repeats this criticism and adds to it the criticism that choices involving risk cannot be rationalized by the ordinal properties of utility functions.

²⁴ For a rigorous presentation of the second statement and a rigorous proof that the statements are equivalent see von Neumann and Morgenstern, *op. cit.*, pp. 26–27, 617–32.

with it, and so on. We can abstract from these factors, which play no role in the present problem, by supposing either that they are the same for different incomes compared or that they can be converted into equivalent sums of money income.²⁵ This permits us to consider total utility a function of money income alone.

Let I represent the income of a consumer unit per unit time, and $U(I)$ the utility attached to that income if it is regarded as certain. Measure I along the horizontal axis of a graph and U along the vertical. In general, $U(I)$ will not be defined for all values of I , since there will be a lower limit to the income a consumer unit can receive, namely, a negative income equal (in absolute value) to the maximum amount that the consumer unit can lose per unit time for the period to which the utility curve refers.

Alternatives open to the consumer unit that involve no risk consist of possible incomes, say I', I'', \dots . The hypothesis then implies simply that the consumer unit will choose the income to which it attaches the most utility. Other things the same, we know from even casual observation that the consumer unit will in general choose the largest income: put differently, we consider it pathological for an individual literally to throw money away, yet this means of choosing a smaller income is always available. It follows that the hypothesis can rationalize riskless choices of the limited kind considered here if, and only if, the utility of money income is larger, the higher the income. Consideration of riskless choices imposes no further requirements on the utility function.

Alternatives involving risk consist of probability distributions of possible incomes. Fortunately, it will suffice for our purpose to consider only a particularly simple kind of alternative involving risk, namely (A) a chance a ($0 < a < 1$) of an income I_1 , and a chance $(1 - a)$ of an income I_2 , where for simplicity I_2 is supposed always greater than I_1 . This simplification is possible because, as we shall see later, the original hypothesis implies that choices of consumer units among more complicated alternatives can be predicted from complete knowledge of their preferences among alternatives like A and a riskless alternative (B) consisting of a certain income I_0 .

Since "other things" are supposed the same for alternatives A and B, the utility of the two alternatives may be taken to be functions solely of the incomes and probabilities involved and not also of attendant

²⁵ The other factors abstracted from must not, of course, include any that cannot in fact be held constant while money income varies. For example, a higher income is desired because it enables a consumer unit to purchase a wider variety of commodities. The consumption pattern of the consumer unit must not therefore be supposed to be the same at different incomes. As another example, a higher income may mean that a consumer unit must pay a higher price for a particular commodity (e.g., medical service). Such variation in price should not be impounded in *ceteris paribus*, though price changes not necessarily associated with changes in the consumer unit's income should be.

circumstances. The utility of alternative B is $U(I_0)$. The expected utility of A is given by

$$\bar{U}(A) = aU(I_1) + (1 - a)U(I_2).$$

According to the hypothesis, a consumer unit will choose A if $\bar{U} > U(I_0)$, will choose B if $\bar{U} < U(I_0)$, and will be indifferent between A and B if $\bar{U} = U(I_0)$.

Let $\bar{I}(A)$ be the actuarial value of A , i.e., $\bar{I}(A) = aI_1 + (1 - a)I_2$. If I_0 is equal to \bar{I} , the "gamble" or "insurance" is said to be "fair" since the consumer unit gets the same actuarial value whichever alternative it chooses. If, under these circumstances, the consumer unit chooses A , it shows a preference for this risk. This is to be interpreted as meaning that $\bar{U} > U(\bar{I})$ and indeed $\bar{U} - U(\bar{I})$ may be taken to measure the utility it attaches to this particular risk.²⁶ If the consumer unit chooses B , it shows a preference for certainty. This is to be interpreted as meaning that $\bar{U} < U(\bar{I})$. Indifference between A and B is to be interpreted as meaning that $\bar{U} = U(\bar{I})$.

Let I^* be the certain income that has the same utility as A , that is, $U(I^*) = \bar{U}$.²⁷ Call I^* the income equivalent to A . The requirement, derived from consideration of riskless choices, that utility increase with income means that

$$\bar{U} \cong U(\bar{I})$$

implies

$$I^* \cong \bar{I}.$$

If I^* is greater than \bar{I} , the consumer unit prefers this particular risk to a certain income of the same actuarial value and would be willing to pay a maximum of $I^* - \bar{I}$ for the privilege of "gambling." If I^* is less than \bar{I} , the consumer unit prefers certainty and is willing to pay a maximum of $\bar{I} - I^*$ for "insurance" against this risk.

These concepts are illustrated for a consumer unit who is willing to pay for insurance ($\bar{I} > I^*$) in Figure 1, *a*, and for a consumer unit who is willing to pay for the privilege of gambling ($\bar{I} < I^*$) in Figure 1, *b*. In

²⁶ This interpretation of $\bar{U} - U(\bar{I})$ as the utility attached to a particular risk is directly relevant to a point to which von Neumann and Morgenstern and commentators on their work have given a good deal of attention, namely, whether there may "not exist in an individual a (positive or negative) utility of the mere act of 'taking a chance,' of gambling, which the use of the mathematical expectation obliterates" (von Neumann and Morgenstern, *op. cit.*, p. 28). In our view the hypothesis is better interpreted as a rather special explanation why gambling has utility or disutility to a consumer unit, and as providing a particular measure of the utility or disutility, than as a denial that gambling has utility (see *ibid.*, pp. 28, 629-32).

²⁷ Since U has been assumed strictly monotonic to rationalize riskless choices, there will be only one income, if any, that has the same utility as A . There will be one if U is continuous which, for simplicity, we assume to be the case throughout this paper.

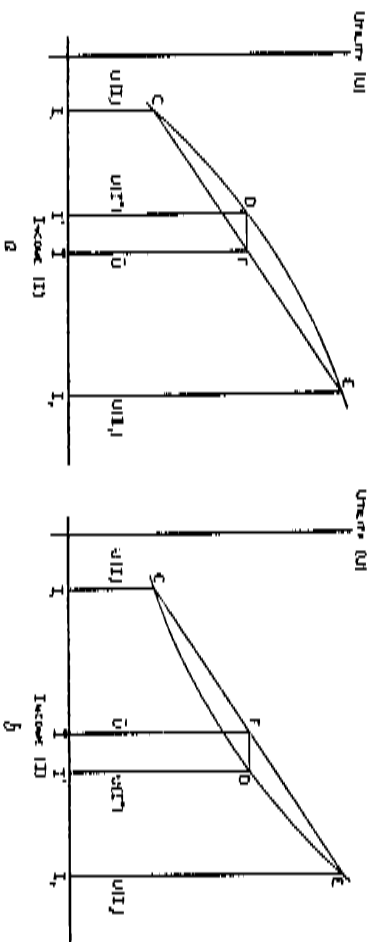


FIG. 1.—Illustration of utility analysis of choices involving risk: a , preference for certainty; b , preference for risk

both figures, money income is measured along the horizontal axis, and utility along the vertical. On the horizontal axis, designate I_1 and I_2 . \bar{I} , the actuarial value of I_1 and I_2 , is then represented by a point that divides the interval I_1 to I_2 in the proportion

$$\frac{1-a}{a} \left(\text{i.e., } \frac{\bar{I} - I_1}{I_2 - \bar{I}} = \frac{1-a}{a} \right).$$

Draw the utility curve (CDE in both figures). Connect the points $(I_1, U[I_1])$, $(I_2, U[I_2])$ by a straight line (CFE). The vertical distance of this line from the horizontal axis at \bar{I} is then equal to \bar{U} . (Since \bar{I} divides the distance between I_1 and I_2 in the proportion $[1-a]/a$, F divides the vertical distance between C and E in the same proportion, so the vertical distance from F to the horizontal axis is the expected value of $U[I_1]$ and $U[I_2]$.) Draw a horizontal line through F and find the income corresponding to its intersection with the utility curve (point D). This is the income the utility of which is the same as the expected utility of A , hence by definition is I^* .

In Figure 1, *a*, the utility curve is so drawn as to make I^* less than \bar{I} . If the consumer unit is offered a choice between A and a certain income I_0 greater than I^* , it will choose the certain income. If this certain income I_0 were less than \bar{I} , the consumer unit would be paying $\bar{I} - I_0$ for certainty—in ordinary parlance it would be “buying insurance”; if the certain income were greater than \bar{I} , it would be being paid $I_0 - \bar{I}$ for accepting certainty, even though it is willing to pay for certainty—we might say that it is “selling a gamble” rather than “buying insurance.” If the consumer unit were offered a choice between A and a certain income I_0 less than I^* , it would choose A because, while it is willing to pay a price for certainty, it is being asked to pay more than the maximum amount ($\bar{I} - I^*$) that it is willing to pay. The price of insurance has become so high that it has, as it were, been converted into a seller rather than a buyer of insurance.

In Figure 1, *b*, the utility curve is so drawn as to make I^* greater than \bar{I} . If the consumer unit is offered a choice between A and a certain income less than I^* , it will choose A . If this certain income I_0 were greater than \bar{I} , the consumer unit would be paying $I_0 - \bar{I}$ for this risk—in ordinary parlance, it would be choosing to gamble or, one might say, “to buy a gamble”; if the certain income were less than \bar{I} , it would be being paid $\bar{I} - I_0$ for accepting this risk even though it is willing to pay for the risk—we might say that it is “selling insurance” rather than “buying a gamble.” If the consumer unit is offered a choice between A and a certain income I_0 greater than I^* , it will choose the certain income because, while it is willing to pay something for a gamble, it is not willing

to pay more than $I^* - \bar{I}$. The price of the gamble has become so high that it is converted into a seller, rather than a buyer, of gambles.

It is clear that the graphical condition for a consumer unit to be willing to pay something for certainty is that the utility function be above its chord at \bar{I} . This is simply a direct translation of the condition that $U(\bar{I}) > \bar{U}$. Similarly, a consumer unit will be willing to pay something for a risk if the utility function is below its chord at \bar{I} .

The relationship between these formalized "insurance" and "gambling" situations and what are ordinarily called insurance and gambling is fairly straightforward. A consumer unit contemplating buying insurance is to be regarded as having a current income of I_2 and as being subject to a chance of losing a sum equal to $I_2 - I_1$, so that if this loss should occur its income would be reduced to I_1 . It can insure against this loss by paying a premium equal to $I_2 - I_0$. The premium, in general, will be larger than $I_2 - \bar{I}$, the "loading" being equal to $\bar{I} - I_0$. Purchase of insurance therefore means accepting the certainty of an income equal to I_0 instead of a pair of alternative incomes having a higher expected value. Similarly, a consumer unit deciding whether to gamble (e.g., to purchase a lottery ticket) can be interpreted as having a current income equal to I_0 . It can have a chance $(1 - a)$ of a gain equal to $I_2 - I_0$ by subjecting itself to a chance a of losing a sum equal to $I_0 - I_1$. If it gambles, the actuarial value of its income is \bar{I} , which in general is less than I_0 . $I_0 - \bar{I}$ is the premium it is paying for the chance to gamble (the "take" of the house, or the "banker's cut").

It should be emphasized that this analysis is all an elaboration of a particular hypothesis about the way consumer units choose among alternatives involving risk. This hypothesis describes the reactions of consumer units in terms of a utility function, unique except for origin and unit of measure, which gives the utility assigned to certain incomes and which has so far been taken for granted. Yet for choices among certain incomes only a trivial characteristic of this function is relevant, namely, that it rises with income. The remaining characteristics of the function are relevant only to choices among alternatives involving risk and can therefore be inferred only from observation of such choices. The precise manner in which these characteristics are implicit in the consumer unit's preferences among alternatives involving risk can be indicated most easily by describing a conceptual experiment for determining the utility function.

Select any two incomes, say \$500 and \$1,000. Assign any arbitrary utilities to these incomes, say 0 utiles and 1 utile, respectively. This corresponds to an arbitrary choice of origin and unit of measure. Select any intermediate income, say \$600. Offer the consumer unit the choice between (A) a chance a of \$500 and $(1 - a)$ of \$1,000 or (B) a certainty

of \$600, varying a until the consumer unit is indifferent between the two (i.e., until $I^* = \$600$). Suppose this indifference value of a is $\frac{2}{5}$. If the hypothesis is correct, it follows that

$$\begin{aligned} U(600) &= \frac{2}{5} U(500) + \frac{3}{5} U(1000) \\ &= \frac{2}{5} 0 + \frac{3}{5} \cdot 1 = \frac{3}{5} = .60. \end{aligned}$$

In this way the utility attached to every income between \$500 and \$1,000 can be determined. To get the utility attached to any income outside the interval \$500 to \$1,000, say \$10,000, offer the consumer unit a choice between (A) a chance a of \$500 and $(1 - a)$ of \$10,000 or (B) a certainty of \$1,000, varying a until the consumer unit is indifferent between the two (i.e., until $I^* = \$1,000$). Suppose this indifference value of a is $\frac{4}{5}$. If the hypothesis is correct, it follows that

$$\frac{4}{5} U(500) + \frac{1}{5} U(10,000) = U(1000),$$

or

$$\frac{4}{5} \cdot 0 + \frac{1}{5} U(10,000) = 1,$$

or

$$U(10,000) = 5.$$

In principle, the possibility of carrying out this experiment, and the reproducibility of the results, would provide a test of the hypothesis. For example, the consistency of behavior assumed by the hypothesis would be contradicted if a repetition of the experiment using two initial incomes other than \$500 and \$1,000 yielded a utility function differing in more than origin and unit of measure from the one initially obtained.

Given a utility function obtained in this way, it is possible, if the hypothesis is correct, to compute the utility attached to (that is, the expected utility of) any set or sets of possible incomes and associated probabilities and thereby to predict which of a number of such sets will be chosen. This is the precise meaning of the statement made toward the beginning of this section that, if the hypothesis were correct, complete knowledge of the preferences of consumer units among alternatives like A and B would make it possible to predict their reactions to any other choices involving risk.

The choices a consumer unit makes that involve risk are typically far

more complicated than the simple choice between *A* and *B* that we have used to elaborate the hypothesis. There are two chief sources of complication: Any particular alternative typically offers an indefinitely large number of possible incomes, and "other things" are generally not the same.

The multiplicity of possible incomes is very general: losses insured against ordinarily have more than one possible value; lotteries ordinarily have more than one prize; the possible income from a particular occupation, investment, or business enterprise may be equal to any of an indefinitely large number of values. A hypothesis that the essence of choices among the degrees of risk involved in such complex alternatives is contained in such simple choices as the choice between *A* and *B* is by no means tautological.

The hypothesis does not, of course, pretend to say anything about how consumer choices will be affected by differences in things other than degree of risk. The significance for our purposes of such differences is rather that they greatly increase the difficulty of getting evidence about reactions to differences in risk alone. Much casual experience, particularly experience bearing on what is ordinarily regarded as gambling, is likely to be misinterpreted, and erroneously regarded as contradictory to the hypothesis, if this difficulty is not explicitly recognized. In much so-called gambling the individual chooses not only to bear risk but also to participate in the mechanics of a game of chance; he buys, that is, a gamble, in our technical sense, and entertainment. We can conceive of separating these two commodities: he could buy entertainment alone by paying admission to participate in a game using valueless chips; he could buy the gamble alone by having an agent play the game of chance for him according to detailed instructions.²⁸ Further, insurance and gambles are often purchased in almost pure form. This is notably true of insurance. It is true also of gambling by the purchase of lottery tickets when the purchaser is not a spectator to the drawing of the winners (e.g., Irish sweepstakes tickets bought in this country or the "numbers" game), and of much stock-market speculation.

An example of behavior that would definitely contradict the assertion, contained in the hypothesis, that the same utility function can be used to explain choices that do and do not involve risk would be willingness by an individual to pay more for a gamble than the maximum amount he could win. In order to explain riskless choices it is necessary to suppose that utility increases with income. It follows that the average

²⁸ It does not, of course, follow that the price an individual is willing to pay for the joint commodity is simply the sum of the prices he is willing to pay for them separately. Indeed, it may well be the possible existence of such a difference that people have in mind when they speak of a "specific utility of gambling."

utility of two incomes can never exceed the utility of the larger income and hence that an individual will never be willing to pay, for example, a dollar for a chance of winning, at most, 99 cents.

More subtle observation would be required to contradict the assertion that the reactions of persons to complicated gambles can be inferred from their reactions to simple gambles. For example, suppose an individual refuses an opportunity to toss a coin for a dollar and also to toss a coin for two dollars but then accepts an opportunity to toss two coins in succession, the first to determine whether the second toss is to be for one dollar or for two dollars. This behavior would definitely contradict the hypothesis. On the hypothesis, the utility of the third gamble is an average of the utility of the first two. His refusal of the first two indicates that each of them has a lower utility than the alternative of not gambling; hence, if the hypothesis were correct, the third should have a lower utility than the same alternative, and he should refuse it.

4. Restrictions on Utility Function Required to Rationalize Observable Behavior

The one restriction imposed on the utility function in the preceding section is that total utility increase with the size of money income. This restriction was imposed to rationalize the first of the facts listed below. We are now ready to see whether the behavior described in section 2 can be rationalized by the hypothesis, and, if so, what additional restrictions this behavior imposes on the utility function. To simplify the task, we shall take as a summary of the essential features of the behavior described in section 2 the following five statements, alleged to be facts: (1) consumer units prefer larger to smaller certain incomes; (2) low-income consumer units buy, or are willing to buy, insurance; (3) low-income consumer units buy, or are willing to buy, lottery tickets; (4) many low-income consumer units buy, or are willing to buy, both insurance and lottery tickets; (5) lotteries typically have more than one prize.

These particular statements are selected not because they are the most important in and of themselves but because they are convenient to handle and the restrictions imposed to rationalize them turn out to be sufficient to rationalize all the behavior described in section 2.

It is obvious from Figure 1 and our discussion of it that if the utility function were everywhere convex from above (for utility functions with a continuous derivative, if the marginal utility of money does not increase for any income), the consumer unit, on our hypothesis, would be willing to enter into any fair insurance plan but would be unwilling to pay anything in excess of the actuarial value for any gamble. If the

utility function were everywhere concave from above (for functions with a continuous derivative, if the marginal utility of money does not diminish for any income), the consumer unit would be willing to enter into any fair gamble but would be unwilling to pay anything in excess of the actuarial value for insurance against any risk.

It follows that our hypothesis can rationalize statement 2, the purchase of insurance by low-income consumer units, only if the utility functions of the corresponding units are not everywhere concave from above; that it can rationalize statement 3, the purchase of lottery tickets by low-income consumer units, only if the utility functions of the corresponding units are not everywhere convex from above; and that it can rationalize statement 4, the purchase of both insurance and lottery tickets by low-income consumer units, only if the utility functions of the corresponding units are neither everywhere concave from above nor everywhere convex from above.

The simplest utility function (with a continuous derivative) that can rationalize all three statements simultaneously is one that has a segment convex from above followed by a segment concave from above and no other segments.²⁹ The convex segment must precede the concave segment because of the kind of insurance and of gambling the low-income consumer units are said to engage in: a chord from the existing income to a lower income must be below the utility function to rationalize the purchase of insurance against the risk of loss; a chord from the immediate neighborhood of the existing income to a higher income must be above the utility function at the existing income to rationalize the purchase for a small sum of a small chance of a large gain.³⁰

Figure 2 illustrates a utility function satisfying these requirements. Let this utility function be for a low-income consumer unit whose current income is in the initial convex segment, say at the point designated I^* . If some risk should arise of incurring a loss, the consumer unit would clearly (on our hypothesis) be willing to insure against the loss (if it did not have to pay too much "loading") since a chord from the utility curve at I^* to the utility curve at the lower income that would be the consequence of the actual occurrence of the loss would everywhere be below the utility function. The consumer unit would not be willing to engage in small gambling. But suppose it is offered a fair gamble of the kind represented by a lottery involving a small chance of winning a relatively

²⁹ A kink or a jump in the utility function could rationalize either the gambling or the insurance. For example, the utility function could be composed of two convex or two concave segments joined in a kink. There is no essential loss in generality in neglecting such cases, as we shall do from here on, since one can always think of rounding the kink ever so slightly.

³⁰ If there are more than two segments and a continuous derivative, a convex segment necessarily precedes a concave segment.

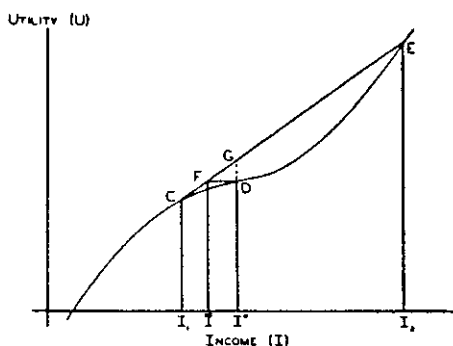


FIG. 2.—Illustration of utility function consistent with willingness of a low-income consumer unit both to purchase insurance and to gamble.

large sum equal to $I_2 - I^*$ and a large chance of losing a relatively small sum equal to $I^* - I_1$. The consumer unit would clearly prefer the gamble, since the expected utility (I^*G) is greater than the utility of I^* . Indeed it would be willing to pay any premium up to $I^* - \bar{I}$ for the privilege of gambling; that is, even if the expected value of the gamble were almost as low as \bar{I} , it would accept the gamble in preference to a certainty of receiving I^* . The utility curve in Figure 2 is therefore clearly consistent with statements 2, 3, and 4.

These statements refer solely to the behavior of relatively low-income consumer units. It is tempting to seek to restrict further the shape of the utility function, and to test the restrictions so far imposed, by appealing to casual observation of the behavior of relatively high-income consumer units.³¹ It does not seem desirable to do so, however, for two major reasons: (1) it is far more difficult to accumulate reliable information about the behavior of relatively high-income consumer units than about the behavior of the more numerous low-income units; (2) perhaps even more important, the progressive income tax so affects the terms under which the relatively high-income consumer units purchase insurance or gamble as to make evidence on their behavior hard to

³¹ For example, a high-income consumer unit that had a utility function like that in Fig. 2 and a current income of I_2 would be willing to participate in a wide variety of gambling, including the purchase of lottery tickets; it would be unwilling to insure against losses that had a small expected value (i.e., involved payment of a small premium) though it might be willing to insure against losses that had a large expected value. Consequently, unwillingness of relatively high-income consumer units to purchase lottery tickets, or willingness to purchase low-premium insurance, would contradict the utility function of Fig. 2 and require the imposition of further restrictions.

interpret for our purposes.³² Therefore, instead of using observations about the behavior of relatively high-income consumer units, we shall seek to learn more about the upper end of the curve by using statement 5, the tendency for lotteries to have more than one prize.

In order to determine the implications of this statement for the utility function, we must investigate briefly the economics of lotteries. Consider an entrepreneur conducting a lottery and seeking to maximize his income from it. For simplicity, suppose that he conducts the lottery by deciding in advance the number of tickets to offer and then auctioning them off at the highest price he can get.³³ Aside from advertising and the like, the variables at his disposal are the terms of the lottery: the number of tickets to sell, the total amount to offer as prizes (which together, of course, determine the actuarial value of a ticket), and the structure of prizes to offer. For any given values of the first two, the optimum structure of prizes is clearly that which maximizes the price he can get per ticket or, what is the same thing, the excess of the price of a ticket over its actuarial value—the “loading” per ticket.

In the discussion of Figure 2, it was noted that $I^* - \bar{I}$ was the maximum amount in excess of the actuarial value that the corresponding consumer unit would pay for a gamble involving a chance $(1 - a)$ of winning $I_2 - I^*$ and a chance a of losing $I^* - I_1$. This gamble is equivalent to a lottery offering a chance $(1 - a)$ of a prize $I_2 - I_1$, in return

³² The effect of the income tax, already referred to in n. 16 above, depends greatly on the specific provisions of the tax law and of the insurance or gambling plan. For example, if an uninsured loss is deductible in computing taxable income (as is loss of an owned home by fire under the federal income tax) while the premium for insuring against the loss is not (as a fire-insurance premium on an owned home is not), the expected value of the loss is less to the consumer unit than to the firm selling insurance. A premium equal to the actuarial value of the loss to the insurance company then exceeds the actuarial value of the loss to the consumer unit. That is, the government in effect pays part of the loss but none of the premium. On the other hand, if the premium is deductible (as a health-insurance premium may be), while an uninsured loss is not (as the excess of medical bills over \$2,500 for a family is not), the net premium to the consumer unit is less than the premium received by the insurance company. Similarly, gambling gains in excess of gambling losses are taxable under the federal income tax, while gambling losses in excess of gambling gains are not deductible. The special treatment of capital gains and losses under the existing United States federal income tax adds still further complications.

Even if both the premium and the uninsured loss are deductible, or a gain taxable and the corresponding loss deductible, the income tax may change the terms because of the progressive rates. The tax-saving from a large loss may be a smaller fraction of the loss than the tax payable on the gain is of the gain.

These comments clearly apply not only to insurance and gambling proper but also to other economic decisions involving risk—the purchase of securities, choice of occupation or business, etc. The neglect of these considerations has frequently led to the erroneous belief that a progressive income tax does not affect the allocation of resources and is in this way fundamentally different from excise taxes.

³³ This was, in fact, the way in which the British government conducted many of its official lotteries. It frequently auctioned off the tickets to lottery dealers, who served as the means of distributing the tickets to the public (see Ewen, *op. cit.*, pp. 234–40).

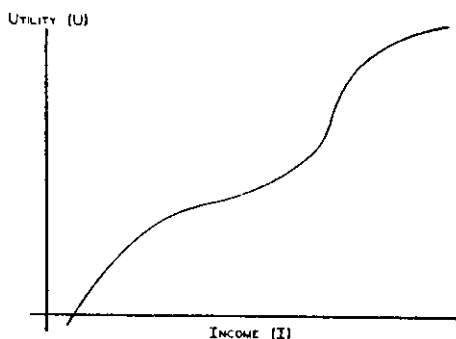


FIG. 3.—Illustration of typical shape of utility curve

for the purchase of a ticket at a price of $I^* - I_1$, the chance of winning the prize being such that $\bar{I} - I_1$ is the actuarial worth of a ticket (i.e., is equal to $[1 - a] \times [I_2 - I_1]$). If the consumer unit won the prize, its net winnings would be $I_2 - I^*$, since it would have to subtract the cost of the ticket from the gross prize. The problem of the entrepreneur, then, is to choose the structure of prizes that will maximize $I^* - \bar{I}$ for a given actuarial value of a ticket, that is, for a given value of $\bar{I} - I_1$. Changes in the structure of prizes involve changes in $I_2 - I_1$. If there is a single prize, $I_2 - I_1$ is equal to the total amount to be distributed ($[1 - a]$ is equal to the reciprocal of the number of tickets). If there are two equal prizes, $I_2 - I_1$ is cut in half ($[1 - a]$ is then equal to twice the reciprocal of the number of tickets). Suppose Figure 2 referred to this latter situation in which there were two equal prizes, I^* on the diagram designating both the current income of the consumer unit and the income equivalent to the lottery. If the price and actuarial worth of the ticket were kept unchanged, but a single prize was substituted for the two prizes (and $[1 - a]$ correspondingly reduced), the gamble would clearly become more attractive to the consumer unit. I_2 would move to the right, the chord connecting $U(I_1)$ and $U(I_2)$ would rotate upward, \bar{U} would increase, and the consumer unit would be paying less than the maximum amount it was willing to pay. The price of the ticket could accordingly be increased; that is, I_2 , \bar{I} and I_1 could be moved to the left until the I^* for the new gamble were equal to the consumer unit's current income (the I^* for the old gamble). The optimum structure of prizes clearly consists therefore of a single prize, since this makes $I_2 - I_1$ as large as possible.

Statement 5, that lotteries typically have more than one prize, is therefore inconsistent with the utility function of Figure 2. This additional fact can be rationalized by terminating the utility curve with a suitable convex segment. This yields a utility curve like that drawn in Figure 3.

With such a utility curve, $I^* - \bar{I}$ would be a maximum at the point at which a chord from $U(I_1)$ was tangent to the utility curve, and a larger prize would yield a smaller value of $I^* - \bar{I}$.³⁴

A utility curve like that drawn in Figure 3 is the simplest one consistent with the five statements listed at the outset of this section.

5. A Digression

It seems well to digress at this point to consider two questions that, while not strictly relevant to our main theme, are likely to occur to many readers: first, is not the hypothesis patently unrealistic; second, can any plausible interpretation be given to the rather peculiar utility function of Figure 3?

a) The Descriptive "Realism" of the Hypothesis

An objection to the hypothesis just presented that is likely to be raised by many, if not most, readers is that it conflicts with the way human beings actually behave and choose. Is it not patently unrealistic to suppose that individuals consult a wiggly utility curve before gambling or buying insurance, that they know the odds involved in the gambles or insurance plans open to them, that they can compute the expected utility of a gamble or insurance plan, and that they base their decision on the size of the expected utility?

While entirely natural and understandable, this objection is not strictly relevant. The hypothesis does not assert that individuals explicitly or consciously calculate and compare expected utilities. Indeed, it is not at all clear what such an assertion would mean or how it could be tested. The hypothesis asserts rather that, in making a particular class of decisions, individuals behave *as if* they calculated and compared expected utility and *as if* they knew the odds. The validity of this assertion does not depend on whether individuals know the precise odds, much less on whether they say that they calculate and compare expected utilities

³⁴ An additional convex segment guarantees that there will always exist current incomes of the consumer unit for which (a) attractive gambles exist and (b) the optimum prize for attractive gambles has a maximum. It does not guarantee that b will be true for every income for which attractive gambles exist. The condition on the current income that attractive gambles exist is that the tangent to the utility curve at the current income be below the utility curve for some income (this argument, like many in later technical footnotes, holds not only for the utility function of Fig. 3 but for any differentiable utility function). A single prize will be the optimum, no matter what the amount distributed in prizes or the fixed actuarial worth of the prize if, and only if, every chord from the utility curve at the current income to the utility of a higher income is everywhere above the utility curve. A particular, and somewhat interesting, class of utility functions for which b will be true for every income for which a is true is the class for which utility approaches a finite limit as income increases.

or think that they do, or whether it appears to others that they do, or whether psychologists can uncover any evidence that they do, but solely on whether it yields sufficiently accurate predictions about the class of decisions with which the hypothesis deals. Stated differently, the test by results is the only possible method of determining whether the *as if* statement is or is not a sufficiently good approximation to reality for the purpose at hand.

A simple example may help to clarify the point at issue. Consider the problem of predicting, before each shot, the direction of travel of a billiard ball hit by an expert billiard player. It would be possible to construct one or more mathematical formulas that would give the directions of travel that would score points and, among these, would indicate the one (or more) that would leave the balls in the best positions. The formulas might, of course, be extremely complicated, since they would necessarily take account of the location of the balls in relation to one another and to the cushions and of the complicated phenomena introduced by "english." Nonetheless, it seems not at all unreasonable that excellent predictions would be yielded by the hypothesis that the billiard player made his shots *as if* he knew the formulas, could estimate accurately by eye the angles, etc., describing the location of the balls, could make lightning calculations from the formulas, and could then make the ball travel in the direction indicated by the formulas. It would in no way disprove or contradict the hypothesis, or weaken our confidence in it, if it should turn out that the billiard player had never studied any branch of mathematics and was utterly incapable of making the necessary calculations: unless he was capable in some way of reaching approximately the same result as that obtained from the formulas, he would not in fact be likely to be an expert billiard player.

The same considerations are relevant to our utility hypothesis. Whatever the psychological mechanism whereby individuals make choices, these choices appear to display some consistency, which can apparently be described by our utility hypothesis. This hypothesis enables predictions to be made about phenomena on which there is not yet reliable evidence. The hypothesis cannot be declared invalid for a particular class of behavior until a prediction about that class proves false. No other test of its validity is decisive.

b) A Possible Interpretation of the Utility Function

A possible interpretation of the utility function of Figure 3 is to regard the two convex segments as corresponding to qualitatively different socioeconomic levels, and the concave segment to the transition between the two levels. On this interpretation, increases in income that raise the relative position of the consumer unit in its own class but do not shift

the unit out of its class yield diminishing marginal utility, while increases that shift it into a new class, that give it a new social and economic status, yield increasing marginal utility. An unskilled worker may prefer the certainty of an income about the same as that of the majority of unskilled workers to an actuarially fair gamble that at best would make him one of the most prosperous unskilled workers and at worst one of the least prosperous. Yet he may jump at an actuarially fair gamble that offers a small chance of lifting him out of the class of unskilled workers and into the "middle" or "upper" class, even though it is far more likely than the preceding gamble to make him one of the least prosperous unskilled workers. Men will and do take great risks to distinguish themselves, even when they know what the risks are. May not the concave segment of the utility curve of Figure 3 translate the economic counterpart of this phenomenon appropriately?

A number of additions to the hypothesis are suggested by this interpretation. In the first place, may there not be more than two qualitatively distinguishable socioeconomic classes? If so, might not each be reflected by a convex segment in the utility function? At the moment, there seems to be no observed behavior that requires the introduction of additional convex segments, so it seems undesirable and unnecessary to complicate the hypothesis further. It may well be, however, that it will be necessary to add such segments to account for behavior revealed by further empirical evidence. In the second place, if different segments of the curve correspond to different socioeconomic classes, should not the dividing points between the segments occur at roughly the same income for different consumer units in the same community? If they did, the fruitfulness of the hypothesis would be greatly extended. Not only could the general shape of the utility function be supposed typical; so also could the actual income separating the various segments. The initial convex segment could be described as applicable to "relatively low-income consumer units" and the terminal convex segment as applicable to "relatively high-income consumer units"; and the groups so designated could be identified by the actual income or wealth of different consumer units.

Interpreting the different segments of the curve as corresponding to different socioeconomic classes would, of course, still permit wide variation among consumer units in the exact shape and height of the curve. In addition, it would not be necessary to suppose anything more than rough similarity in the location of the incomes separating the various segments. Different socioeconomic classes are not sharply demarcated from one another; each merges into the next by imperceptible gradations (which, of course, accounts for the income range encompassed by the concave segment); and the generally accepted dividing line between classes will vary from time to time, place to place, and consumer unit to consumer unit. Finally, it is not necessary that every consumer

unit have a utility curve like that in Figure 3. Some may be inveterate gamblers; others, inveterately cautious. It is enough that many consumer units have such a utility curve.

6. Further Implications of Hypothesis

To return to our main theme, we have two tasks yet to perform: first, to show that the utility function of Figure 3 is consistent with those features of the behavior described in section 2 not used in deriving it; second, to suggest additional implications of the hypothesis capable of providing a test of it.

The chief generalization of section 2 not so far used is that people must in general be paid a premium to induce them to bear moderate risks instead of either small or large risks. Is this generalization consistent with the utility function of Figure 3?

It clearly is for a consumer unit whose income places it in the initial convex segment. Such a relatively low-income consumer unit will be willing to pay something more than the actuarial value for insurance against any kind of risk that may arise; it will be averse to small fair gambles; it may be averse to all fair gambles; if not, it will be attracted by fair gambles that offer a small chance of a large gain; the attractiveness of such gambles, with a given possible loss and actuarial value, will initially increase as the size of the possible gain increases and will eventually decrease.³⁵ Such consumer units therefore prefer either certainty

³⁵ The willingness of a consumer unit in the initial convex segment to pay something more than the actuarial value for insurance against any kind of risk follows from the fact that a chord connecting the utility of its current income with the utility of any lower income to which it might be reduced by the risk in question will everywhere be below the utility curve. The expected utility is therefore less than the utility of the expected income.

To analyze the reaction of such a consumer unit to different gambles, consider the limiting case in which the gamble is fair, i.e., $\bar{I} = I_0$. \bar{I} then is both the expected income of the consumer unit if it takes the gamble and its actual income if it does not (i.e., its current income). The possible gains (and associated probabilities) that will be attractive to the unit for a given value of I_1 (i.e., a given possible loss) can be determined by drawing a straight line through $U(I_1)$ and $U(\bar{I})$. All values of $I_2 > \bar{I}$ for which $U(I_2)$ is greater than the ordinate of the extended straight line will be attractive; no others will be.

Since \bar{I} is assumed to be in the first convex segment, there will always exist some values of $I_2 > \bar{I}$ for which $U(I_2)$ is less than the ordinate of the extended straight line. This is the basis for the statement that the consumer unit will be averse to small gambles.

Consider the line that touches the curve at only two points and is nowhere below the utility curve. Call the income at the first of the points at which it touches the curve, which may be the lowest possible income, I' , and the income at the second point, I'' . The consumer unit will be averse to all gambles if its income ($I_0 = \bar{I}$) is equal to or less than I' . This follows from the fact that a tangent to the curve at \bar{I} will then be steeper than the "double tangent" and will intersect the latter prior to I' ; a chord from \bar{I} to a lower income will be even steeper. This is the basis for the statement that the consumer unit may be averse to all gambles.

If the income is above I' , there will always be some attractive gambles. These will offer

or a risk that offers a small chance of a large gain to a risk that offers the possibility of moderate gains or losses. They will therefore have to be paid a premium to induce them to undertake such moderate risks.

The generalization is clearly false for a consumer unit whose income places it in the concave segment. Such an "intermediate-income" consumer unit will be attracted by every small fair gamble; it may be attracted by every fair gamble; it may be averse to all fair insurance; if not, it will be attracted by insurance against relatively large losses.³⁶ Such consumer units will therefore be willing to pay a premium in order to assume moderate risks.

The generalization is partly true, partly false, for a consumer unit whose income places it in the terminal convex segment. Such a relatively high-income consumer unit will be willing to insure against any small possible loss and may be attracted to every fair insurance plan; the only insurance plans it may be averse to are plans involving rather large losses; it may be averse to all fair gambles; if not, it will be attracted by gambles that involve a reasonably sure, though fairly small, gain, with a small possibility of a sizable loss; it will be averse to gambles of the lottery variety.³⁷ These consumer units therefore prefer certainty to moderate risks; in this respect they conform to the generalization. However, they may prefer moderate risks to extreme risks, though these adjectives hardly suffice to characterize the rather complex pattern of risk preferences implied for high-income consumer units by a utility curve like that of Figure 3. Nonetheless, in this respect the implied behavior of the high-income consumer units is either neutral or contrary to the generalization.

Our hypothesis does not therefore lead inevitably to a rate of return higher to uses of resources involving moderate risk than to uses involving little or much risk. It leads to a rate of return higher for uses involving moderate risk than for uses involving little risk only if consumer units in the two convex segments outweigh in importance, for the resource

a small chance of a large gain. The statement about the changing attractiveness of the gamble as the size of the possible gain changes follows from the analysis in sec. 4 of the conditions under which it would be advantageous to have a single prize in a lottery.

³⁶ Consider the tangent to the utility curve at the income the consumer unit would have if it did not take the gamble ($\bar{I} = I_0$). If this income is in the concave section, the tangent will be below the utility curve at least for an interval of incomes surrounding \bar{I} . A chord connecting any two points of the utility curve on opposite sides of \bar{I} and within this interval will always be above the utility curve at \bar{I} (i.e., the expected utility will be above the utility of the expected income), so these gambles will be attractive. The tangent may lie below the utility curve for all incomes. In this case, every fair gamble will be attractive. The unit will be averse to insuring against a loss, whatever the chance of its occurring, if a chord from the current income to the lower income to which it would be reduced by the loss is everywhere above the utility curve. This will surely be true for small losses and may be true for all possible losses.

³⁷ These statements follow directly from considerations like those in the two preceding footnotes.

use in question, consumer units in the concave segment.³⁸ Similarly, it leads to a rate of return higher for uses involving moderate risk than for uses involving much risk only if consumer units in the initial convex segment outweigh in importance consumer units in both the concave and the terminal convex segments—though this may be a more stringent condition than is necessary in view of the uncertainty about the exact role of consumer units in the terminal convex segment.

This relative distribution of consumer units among the various segments could be considered an additional restriction that would have to be imposed to rationalize the alleged higher rate of return to moderately risky uses of resources. It is not clear, however, that it need be so considered, since there are two independent lines of reasoning that, taken together, establish something of a presumption that relatively few consumer units are in the concave segment.

One line of reasoning is based on the interpretation of the utility function suggested in section 5*b* above. If the concave segment is a border line between two qualitatively different social classes, one would expect relatively few consumer units to be between the two classes.

The other line of reasoning is based on the implications of the hypothesis for the relative stability of the economic status of consumer units in the different segments. Units in the intermediate segment are tempted by every small gamble and at least some large ones. If opportunities are available, they will be continually subjecting themselves to risk. In consequence, they are likely to move out of the segment; upwards, if they are lucky; downwards, if they are not. Consumer units in the two convex segments, on the other hand, are less likely to move into the intermediate segment. The gambles that units in the initial segment accept will rarely pay off and, when they do, are likely to shift them all the way into the terminal convex segment. The gambles that units in the terminal segment accept will rarely involve losses and, when they do, may shift them all the way into the lower segment. Under these conditions, maintenance of a stable distribution of the population among the three segments would require that the two convex segments contain many more individuals than the concave segment. These considerations, while persuasive, are not, of course, conclusive. Opportunities to assume risks may not exist. More important, the status of consumer units is determined not alone by the outcome of risks deliberately assumed but also by random events over which they cannot choose and have no control; and it is conceivable that these random events might

³⁸ This statement is deliberately vague. The actual relative rates of return will depend not only on the conditions of demand for risks of different kinds but also on the conditions of supply, and both would have to be taken into account in a comprehensive statement.

be distributed in such a way that their main effect was to multiply the number in the concave segment.

The absolute number of persons in the various segments will count most for choices among the uses of human resources; wealth will count most for choices among uses of nonhuman resources.³⁹ In consequence, one might expect that the premium for bearing moderate risks instead of large risks would be greater for occupations than for investments. Indeed, for investments, the differential might in some cases be reversed, since the relatively high-income consumer units (those in the terminal segment) count for more in wealth than in numbers and they may prefer moderate to extreme risks.

In judging the implications of our hypothesis for the market as a whole, we have found it necessary to consider separately its implications for different income groups. These offer additional possibilities of empirical test. Perhaps the most fruitful source of data would be the investment policies of different income groups.

It was noted in section 2 that, although many persons with low incomes are apparently willing to buy extremely speculative stocks, the low-income group receives the bulk of its property income in the form of interest and rents. These observations are clearly consistent with our hypothesis. Relatively high-income groups might be expected, on our hypothesis, to prefer bonds and relatively safe stocks. They might be expected to avoid the more speculative common stocks but to be attracted to higher-grade preferred stocks, which pay a higher nominal rate of return than high-grade bonds to compensate for a small risk of capital loss. Intermediate income groups might be expected to hold relatively large shares of their assets in moderately speculative common stocks and to furnish a disproportionate fraction of entrepreneurs.

Of course, any empirical study along these lines will have to take into account, as noted above, the effect of the progressive income tax in modifying the terms of investment. The current United States federal income tax has conflicting effects: the progressive rates discourage risky investments; the favored treatment of capital gains encourages them. In addition, such a study will have to consider the risk of investments as a group, rather than of individual investments, since the rich may be in a position to "average" risks.

Another implication referred to above that may be susceptible of empirical test, and the last one we shall cite, is the implied difference in the stability of the relative income status of various economic groups. The unattractiveness of small risks to both high- and low-income consumer

³⁹ This distinction requires qualification because of the need for capital to enter some types of occupations and the consequent existence of "noncompeting groups"; see Milton Friedman and Simon Kuznets, *Income from Independent Professional Practice* (New York: National Bureau of Economic Research, 1945), chap. iii, sec. 3; chap. iv, sec. 2.

units would tend to give them a relatively stable status. By contrast, suppose the utility curve had no terminal convex segment but was like the curve of Figure 2. Low-income consumer units would still have a relatively stable status: their willingness to take gambles at long odds would pay off too seldom to shift many from one class to another. High-income consumer units would not. They would then take almost any gamble, and those who had high incomes today almost certainly would not have high incomes tomorrow. The average period from “shirt sleeves to shirt sleeves” would be far shorter than “three generations.”⁴⁰ Unlike the other two groups, the middle-income class might be expected to display considerable instability of relative income status.⁴¹

7. Conclusion

A plausible generalization of the available empirical evidence on the behavior of consumer units in choosing among alternatives open to them is provided by the hypothesis that a consumer unit (generally a family, sometimes an individual) behaves as if

1. It had a consistent set of preferences;
2. These preferences could be completely described by attaching a numerical value—to be designated “utility”—to alternatives each of which is regarded as certain;
3. The consumer unit chose among alternatives not involving risk that one which has the largest utility;
4. It chose among alternatives involving risk that one for which the expected utility (as contrasted with the utility of the expected income) is largest;
5. The function describing the utility of money income had in general the following properties:
 - a) Utility rises with income, i.e., marginal utility of money income everywhere positive;

⁴⁰ We did not use the absence of such instability to derive the upper convex segment because of the difficulty of allowing for the effect of the income tax.

⁴¹ The existing data on stability of relative income status are too meager to contradict or to confirm this implication. In their study of professional incomes Friedman and Kuznets found that relative income status was about equally stable at all income levels. However, this study is hardly relevant, since it was for homogeneous occupational groups that would tend to fall in a single one of the classes considered here. Mendershausen's analysis along similar lines for family incomes in 1929 and 1933 is inconclusive. See Friedman and Kuznets, *op. cit.*, chap. vii; Horst Mendershausen, *Changes in Income Distribution during the Great Depression* (New York: National Bureau of Economic Research, 1946), chap. iii.

- b) It is convex from above below some income, concave between that income and some larger income, and convex for all higher incomes, i.e., diminishing marginal utility of money income for incomes below some income, increasing marginal utility of money income for incomes between that income and some larger income, and diminishing marginal utility of money income for all higher incomes;
6. Most consumer units tend to have incomes that place them in the segments of the utility function for which marginal utility of money income diminishes.

Points 1, 2, 3, and 5*a* of this hypothesis are implicit in the orthodox theory of choice; point 4 is an ancient idea recently revived and given new content by von Neumann and Morgenstern; and points 5*b* and 6 are the consequence of the attempt in this paper to use this idea to rationalize existing knowledge about the choices people make among alternatives involving risk.

Point 5*b* is inferred from the following phenomena: (*a*) low-income consumer units buy, or are willing to buy, insurance; (*b*) low-income consumer units buy, or are willing to buy, lottery tickets; (*c*) many consumer units buy, or are willing to buy, both insurance and lottery tickets; (*d*) lotteries typically have more than one prize. These statements are taken as a summary of the essential features of observed behavior not because they are the most important features in and of themselves but because they are convenient to handle and the restrictions imposed to rationalize them turn out to be sufficient to rationalize all the behavior described in section 2 of this paper.

A possible interpretation of the various segments of the utility curve specified in 5*b* is that the segments of diminishing marginal utility correspond to socioeconomic classes, the segment of increasing marginal utility to a transitional stage between a lower and a higher socioeconomic class. On this interpretation the boundaries of the segments should be roughly similar for different people in the same community; and this is one of several independent lines of reasoning leading to point 6.

This hypothesis has implications for behavior, in addition to those used in deriving it, that are capable of being contradicted by observable data. In particular, the fundamental supposition that a single utility curve can generalize both riskless choices and choices involving risk would be contradicted if (*a*) individuals were observed to choose the larger of two certain incomes offered to them but (*b*) individuals were willing to pay more than the largest possible gain for the privilege of bearing risk. The supposition that individuals seek to maximize expected utility would be contradicted if individuals' reactions to complicated gambles could not be inferred from their reactions to simple ones. The particular shape

of the utility curve specified in 5*b* would be contradicted by any of a large number of observations, for example, (*a*) general willingness of individuals, whatever their income, who buy insurance against small risks to enter into small fair gambles under circumstances under which they are not also buying "entertainment," (*b*) the converse of *a*, namely an unwillingness to engage in small fair gambles by individuals who are not willing to buy fair insurance against small risks, (*c*) a higher average rate of return to uses of resources involving little risk than to uses involving a moderate amount of risk when other things are the same, (*d*) a concentration of investment portfolios of relatively low-income groups on speculative (but not highly speculative) investments or of relatively high-income groups on either moderately or highly speculative investments, (*e*) great instability in the relative income status of high-income groups or of low-income groups as a consequence of a propensity to engage in speculative activities.

The Expected-Utility Hypothesis and the Measurability of Utility

with L. J. Savage

Renewed attention is currently being devoted to a hypothesis about choices involving risk suggested by Gabriel Cramer and Daniel Bernoulli, employed by Alfred Marshall, and recently revived by von Neumann and Morgenstern: namely, that individuals choose in such circumstances as if they were seeking to maximize the expected value of some quantity. The hypothetical quantity thus defined has, especially recently, been called “utility.” This hypothesis, if reasonably valid in a sufficiently wide domain, has far-reaching implications for economic theory.¹ It provides a unified interpretation of two kinds of economic behavior that have traditionally been rationalized on divergent, and largely inconsistent, lines—first, choices among alternatives regarded as certain, rationalized in terms of consistent preferences for the goods in question and deliberative selection of the alternative highest in the scale of preference; second, choices among alternatives involving risk, rationalized in terms of ill-defined preferences for “risk” or “uncertainty,” generally regarded as “irrational” in the sense that they admit of no simple and reasonably universal description but can only be determined *ad hoc* in each individual case. Furthermore, the expected-utility hypothesis has

We are indebted to William J. Baumol and Jacob Marschak for helpful comments on an earlier draft of this paper.

This paper was mainly written while Savage was a fellow of the John Simon Guggenheim Memorial Foundation and a research scholar in France under the Fulbright Act (Public Law 584, 79th Cong.). His contribution is an outgrowth of work at the University of Chicago sponsored by the Office of Naval Research.

¹ The maximization of such an expected value may also be regarded as a maxim for behavior. As such, it promises to have important implications for statistical theory. Cf. L. J. Savage, “The Theory of Statistical Decision,” *Journal of the American Statistical Association*, XLVI (March, 1951), 55–67. Success of the maxim in this domain, as in any other area of decisions involving uncertainty, depends, not on its empirical verification for the economic behavior of men at large, but on its acceptability, to individuals who are particularly concerned with such decisions, as a rule guiding “wise” behavior in the face of uncertainty.

potentially rich empirical content. As we have shown elsewhere, a few simple and widely accepted empirical generalizations about behavior under circumstances involving risk can be used to specialize the hypothesis sufficiently for it to yield significant implications for phenomena that lend themselves much less readily to casual empirical observation.² These predictions, in turn, when and if verified, will be both a fruit of the hypothesis and a further test of its validity.

Unfortunately, general understanding of the meaning, relevance, and role of the hypothesis has been greatly obstructed by a twofold error in its earlier use: the belief, first, that a numerical utility, unique except for origin and unit of scale, was necessary for the analysis of choices among alternatives regarded as certain; second, that the expectation of this measurable utility must in the same way govern choices among alternatives involving risk. The discovery by Pareto, Slutsky, and later writers that such a numerical utility is not necessary for analyzing riskless choices was correctly regarded as an important theoretical advance. But, in presence of the idea that a utility measure derived from riskless choices must also govern choices among alternatives involving risk, it diverted attention from the possibility that there might be some function, not derivable from riskless choices, the expected value of which governed choices among alternatives involving risk.

In the light of this background, the writers who have recently revived and reinterpreted the hypothesis should have taken special pains to be unmistakably clear about the sense in which the hypothesis justifies or permits the assignment of a measure unique except for origin and unit of scale to preferences for goods and services and perhaps should have used a new name for this measure rather than the name "utility," with all its unfortunate connotations. Unfortunately, they have done neither the one nor the other. Von Neumann and Morgenstern, in reviving the hypothesis and presenting their enlightening axiomatization of it, failed to guard adequately against misunderstanding and, indeed, by some unguarded statements, probably promoted misunderstanding. Similarly, our own exposition of the relation between the hypothesis and available empirical evidence is inadequate in failing to deal directly with the problem of "measurability"; experience demonstrates that our parenthetical allusions to the problem were not sufficiently clear and that they too may have promoted rather than alleviated misunderstanding.

William Baumol's recent note in this *Journal* provides an excellent occasion for seeking to repair in some measure our earlier failure by

² Milton Friedman and L. J. Savage, "The Utility Analysis of Choices Involving Risk," *Journal of Political Economy*, LVI (August, 1948), 279–304; reprinted in American Economic Association, *Readings in Price Theory* (Chicago: Irwin, 1952), pp. 57–96. Experimental use of the hypothesis is reported in a paper by F. Mosteller and P. Nogee, "An Experimental Measurement of Utility," *Journal of Political Economy*, LIX (October, 1951), 371–404.

an attempt to make clear the grounds on which the acceptance or rejection of the hypothesis must rest and the sense in which its acceptance justifies or permits the treatment of utility as “measurable.”³ Baumol’s note is devoted precisely to these questions and states clearly the criticisms of the hypothesis that arise naturally out of the earlier discussions of “measurability.”

Baumol objects to the hypothesis on two grounds: (1) that “the Neumann-Morgenstern utility construction may be incompatible with the scale of preferences of the individual” and (2) that “a Neumann-Morgenstern utility index” may be “unjustifiably arbitrary.”⁴ The first objection raises the general question of the grounds on which the hypothesis should be accepted or rejected (sec. 1); the second, the sense in which the hypothesis “makes” utility “measurable” (sec. 2).

1. Grounds for Accepting or Rejecting the Hypothesis

The function of a scientific hypothesis is to enable us to “predict” phenomena not yet observed, that is, to make statements about phenomena not yet observed that are (a) capable of being contradicted and (b) will not in fact be contradicted. If the statements about unobserved phenomena are not capable of being contradicted, the hypothesis is empty and hence useless for prediction—the hypothesis that a person will choose what he chooses is clearly “correct” and equally clearly “empty,” because incapable of contradiction. If the statements about unobserved phenomena are contradicted (“frequently” or more often and more flagrantly than statements suggested by an alternative hypothesis), then the hypothesis is wrong. The wider the range of observable phenomena capable of contradicting the hypothesis, the greater its *potential* fruitfulness, because this is equivalent to greater precision of prediction. The wider the range of occasions on which the hypothesis has successfully withstood the test of contradiction, the greater the confidence in its validity. The desideratum is a hypothesis that yields highly precise

³ William J. Baumol, “The Neumann-Morgenstern Utility Index—an Ordinalist View,” *Journal of Political Economy*, LIX (February, 1951), 61–66. At one point, Baumol raises the question whether Bernoulli and Marshall maintained the expected-utility hypothesis or the hypothesis that “consumers always accept the actuarial value of a risky proposition as indicative of its utility” (p. 62). This question can be answered definitively. The main thesis of Bernoulli’s paper is precisely the necessity of distinguishing between the two hypotheses to solve the St. Petersburg and similar paradoxes. In the original Latin text, he labels what has come to be called expected utility, “*emolumentum medium*.” Cramer, in the passages from his letter in French to Nicholas Bernoulli that Daniel Bernoulli quotes at the end of his own paper, contrasts the “*espérance mathématique*” with the “*espérance morale*.” Alfred Marshall deals explicitly with expected utility, not expected money value, in his Mathematical Note ix. See Friedman and Savage, *op. cit.*, pp. 280–81, and the references there cited.

⁴ *Op. cit.*, pp. 62 and 65.

predictions (i.e., predictions easily capable of being wrong); that has been used to make many such predictions (i.e., has had many opportunities for contradiction); and that has yielded solely "correct" predictions (i.e., has repeatedly failed to be contradicted).

These trite and oversimplified observations on the basis for choosing among hypotheses are necessary here because their implicit rejection is at the bottom of much of the criticism leveled against the expected-utility hypothesis: the hypothesis is objected to because it is not "empty," because it can be wrong. For example, Baumol says that he can conceive of behavior that would contradict the expected-utility hypothesis and yet would not be clearly "pathological" and that some of the results of the hypothesis "are not introspectively obvious."⁵ In part, Baumol is saying that casual observation and introspection suggest that the hypothesis is false. This assertion is relevant evidence against the hypothesis, since it amounts to saying that crudely observed experience contradicts predictions made by the hypothesis. But it is clear from the examples and context that Baumol is also saying something quite different, which he regards as equally relevant evidence against the hypothesis; namely, that the hypothesis is not obviously and inevitably true, that behavior is conceivable that would contradict it, rather than that such behavior has been observed, however crudely. In light of the preceding paragraph, this feature is clearly a virtue of a scientific hypothesis, not a defect—it is a valid objection only to claims that the hypothesis must be true (i.e., is a tautology). The possibility of specifying behavior that would contradict the hypothesis means that the hypothesis is not empty; that such behavior is not clearly "pathological" or "introspectively obvious" means that the hypothesis, *if valid*, would enable us to make more precise predictions than we can without it.⁶

But granted that the expected-utility hypothesis is *potentially* fruitful and that objections to its acceptance like those just cited are largely

⁵ *Ibid.*, pp. 64 and 65.

⁶ Similarly, Baumol remarks approvingly of a construction of his own that "no situation can have its 'utility' inferred from that of others" (*ibid.*, p. 66). But this simply means that his hypothesis is almost empty and hence almost incapable of adding to our ability to predict behavior; it is therefore of almost no interest at all.

On Baumol's own implicit criterion, ordinal utility analysis of the usual indifference curve variety would have to be rejected. The usual formulations are almost but not quite empty, since they imply at least transitivity. Conceivably someone might choose oranges in preference to apples, and apples in preference to nuts, but nuts in preference to oranges, which would be inconsistent with transitivity.

An analogy with physical science may help clarify the point at issue. There is nothing "pathological" about the attraction between objects varying inversely as the cube of the distance between them, nor is it introspectively obvious that the attraction should vary inversely as the square of the distance. One can readily conceive of a sequence of measurements that if observed would contradict the inverse square law. Clearly none of these statements represents a valid objection to the inverse square law; they merely mean that it has potential power.

irrelevant, what positive reasons are there to accept it? Has it repeatedly survived the test of contradictability? Are there any convincing reasons for believing that it will do so or continue to do so? A highly qualified affirmative answer can be given to both these questions, but no more; as things now stand, the hypothesis, despite its respectable age, must be regarded as a promising conjecture deriving its plausibility more from indirect evidence than from direct survival of "critical" experiments.

Much of the readily available evidence on the consistency of the hypothesis with experience, that is, on the failure of its implications to be contradicted by such experience, is summarized in our earlier article, where some of it is used to specialize the hypothesis. We suggested in that article additional possibilities of contradiction, but so far as we know these have not yet been exploited. Some recent experiments by Mosteller and Nogee⁷ add to the direct observational evidence and fail to contradict the hypothesis. Yet the evidence of this relatively direct kind now available is far from adequate to justify any real confidence in the validity of the hypothesis; at best, it justifies mild optimism.

The very real appeal of the hypothesis derives, we believe, less from this direct evidence than from indirect evidence rendering it plausible that the hypothesis will continue to fail to be contradicted, at least in some important domains. This indirect evidence is provided in some measure by the coherence of the hypothesis with the rest of economic theory. In much greater measure, it is provided by the plausibility of a set of postulates that are sufficient for the derivation of the hypothesis and are themselves derivable from it and so are an alternative statement of the hypothesis. In saying that these postulates are more plausible than the hypothesis to which they are logically equivalent, we mean that the postulates immediately call to mind a host of implications (or predictions from the hypothesis) susceptible to casual empirical observation. With respect to the class of phenomena to which these implications relate, the hypothesis has had many opportunities for contradiction and has repeatedly failed to be contradicted. The evidence is "indirect," because this is not the class of phenomena we are primarily interested in using the hypothesis to predict. Success of the hypothesis for this class makes it plausible that it will succeed for another, not unrelated, class; it does not, however, justify the same confidence as direct evidence for the latter class.

The important original contribution of von Neumann and Morgenstern is precisely to have provided this indirect evidence through their axiomatization of the hypothesis. Since there is a mistake (adverted to by Baumol in his footnote 16 and called to our attention by Paul A.

⁷ *Op. cit.*

Samuelson) in the translation of the Neumann-Morgenstern postulates we presented in our earlier paper,⁸ and since one of us (Savage) has been doing further work on the postulational base of the hypothesis in another connection, a restatement of the postulational evidence is presented here.⁹

Let us suppose, for simplicity, that the final outcome for an individual of a process of choice is among one of some finite number of possibilities, say the set X with elements x_1, x_2, \dots, x_n . These outcomes, the x_i , can be described as the "incomes" that might possibly accrue to the individual, where income is to be understood in the widest sense, embracing, for example, cash income, schedule of cash income over time, baskets of goods, fate in love and war, etc. The assumption that there are only a finite number of outcomes is obviously without any practical significance; mathematically, it is a blemish that can be removed.

In advance, the alternatives available to the individual consist of a set of probability distributions. Any one alternative, frequently called a "prospect," say, the probability distribution f , consists of a probability f_1 that x_1 will be the final outcome, f_2 that x_2 will be, and so on. It can be interpreted as a gamble or lottery ticket offering the individual the probability f_i of receiving the income x_i for each i .

The postulates refer to the set F of all conceivable probability distributions f, g, h, \dots , on the fixed finite set X defined above. Not all of these, of course, need be available to the individual in any particular case. An element f of F may be expressed in the notation $f = [f_1, f_2, \dots, f_n]$, where the f_i 's are nonnegative numbers adding up to unity.

The expression $f \leq g$ is to be read and interpreted, "The gamble f is not preferred (by the individual) to g ," by which is meant that if f and g are the only alternatives available to the individual, he will not systematically choose f .

Though no postulate has yet been introduced, a very important assumption is implicit in the structure thus far defined; namely, that preferences of the person in uncertain situations to which probability applies are governed solely by the probabilities attached to each possible income. This assumption admits and is worthy of some analysis. But this analysis will be passed over here, because we are under the impression that the assumption is, with the usual grain of salt, acceptable to present critics of the utility hypothesis.

⁸ This mistake is noted and corrected in the version reprinted in *Readings in Price Theory*, p. 71.

⁹ The postulate system and the motivation for it presented here may be compared with the corresponding material in Section 3 of J. Marschak, "Why 'Should' Statisticians and Businessmen Maximize 'Moral Expectation'?" *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (Berkeley: University of California Press, 1951), pp. 493-506, with which it is in substantial agreement.

Statement of the postulates is facilitated by the notational convention that if $f = [f_1, \dots, f_n]$, $g = [g_1, \dots, g_n]$, and $0 \leq a \leq 1$; then $af + (1 - a)g = [af_1 + (1 - a)g_1, \dots, af_n + (1 - a)g_n]$. Clearly, $af + (1 - a)g$ is itself an element of F and reduces to f or g when a is 1 or 0, respectively. Quite formally the postulates to be discussed are these.

P1. For all f, g, h (not necessarily distinct) in F :

1. $f \leq g$, or $g \leq f$.

2. If $f \leq g$ and $g \leq h$, then $f \leq h$.

P2. If $af + (1 - a)h \leq g$ for all a such that $0 \leq a < 1$, then $f \leq g$.

P3. For $0 < a < 1$, $af + (1 - a)h \leq ag + (1 - a)h$, if and only if $f \leq g$.

In the presence of P1–3 the utility hypothesis is rigorously implied as a theorem, namely:

THEOREM. There are numbers c_1, \dots, c_n such that $f \leq g$, if and only if

$$\sum f_i c_i \leq \sum g_i c_i$$

Moreover, any two such sequences of numbers c_i and c'_i are connected by an equation

$$c'_i = s + tc_i$$

for some s, t , with $t > 0$.

The proof of this theorem, which is the expected-utility hypothesis, is easy but not appropriate for presentation here.¹⁰ Conversely, it is almost obvious that even the first part of the theorem implies P1–3, verifying an assertion made earlier.

The important question for us is whether the postulates can be expected to correspond reasonably well with observable economic behavior, and this question will now be taken up for the three postulates one by one.

The first postulate states that the individual can be supposed to have a complete and consistent (transitive) ordering of all possible alternatives; that is, one can tell which of two objects the individual prefers or whether he is indifferent between them; and if he does not prefer f to g , and does not prefer g to h , then he does not prefer f to h . These and closely related assumptions have been given much attention by economists. There is widespread agreement that they are introspectively very appealing and that their agreement with experience, though not perfect, is quite good enough to merit continued interest.¹¹

¹⁰ Rather more is proved in the Appendix of John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (2d ed.; Princeton: Princeton University Press, 1947).

¹¹ See Friedman and Savage, *op. cit.*, p. 288, n. 23, and the references therein to the dissenting view by W. E. Armstrong.

The second postulate is a technical assumption of continuity, which seems quite acceptable, though it is not altogether without content. It says, for example, that if a person will not cross the street, no matter how light the traffic, he would not cross even if there were no danger at all.

It is essentially the third postulate that critics of the utility hypothesis seem to find inadequately motivated. But we shall show that this postulate is implied by a principle that we believe practically unique among maxims for wise action in the face of uncertainty, in the strength of its intuitive appeal. The principle is universally known and recognized; and the Greeks must surely have had a name for it, though current English seems not to.

To illustrate the principle before defining it, suppose a physician now knows that his patient has one of several diseases for each of which the physician would prescribe immediate bed rest. We assert that under this circumstance the physician should and, unless confused, will prescribe immediate bed rest whether he is now, later, or never, able to make an exact diagnosis.

Much more abstractly, consider a person constrained to choose between a pair of alternatives, a and b , without knowing whether a particular event E does (or will) in fact obtain. Suppose that, depending on his choice and whether E does obtain, he is to receive one of four (not necessarily distinct) gambles, according to the following schedule.

Choice	Event	
	E	not E
a	$f(a)$	$g(a)$
b	$f(b)$	$g(b)$

The principle in sufficient generality for the present purpose asserts: If the person does not prefer $f(a)$ to $f(b)$, and does not prefer $g(a)$ to $g(b)$, then he will not prefer the choice a to b . Further, if the person does not prefer a to b , he will either not prefer $f(a)$ to $f(b)$ or not prefer $g(a)$ to $g(b)$ (possibly both).

We anticipate that if the reader considers this principle, in the light of the illustration that precedes and such others as he himself may invent, he will concede that the principle is not one he would deliberately violate. This in turn we consider to be some reason for supposing that people do actually tend to avoid flagrant violation of the principle.

In the presence of the assumption (already mentioned) that incomes and probabilities alone determine the preference of a given person, P3 is a consequence of the principle now under discussion. To see this, it is necessary only to set $f(a) = f, f(b) = g, g(a) = g(b) = h$, and to sup-

pose that the probability of E is a . After this specialization, a is the gamble $af + (1 - a)h$, b is $ag + (1 - a)h$, and the principle simply states P3.

2. The “Measurability” of Utility

Baumol objects to the notion that the expected-utility hypothesis enables “the true measure” of utility to be “deduced from . . . the observed behavior of an individual.”¹² There is a sense in which this objection is entirely valid. The significance of the objection is, however, rather different from that implied by Baumol. Since this is an issue on which there is currently much confusion and much writing at cross-purposes, it will be well to start from first principles.

Consider a situation in which an individual must choose one among a set of “prospects.” An extremely general theory of choice is essentially P1 alone; more precisely, that there exists a consistent and transitive ordering of all possible prospects which has the property that the individual will choose the alternative available to him that is highest in this ordering. Let F stand for the set of all possible prospects f, g, \dots ; and $U(f)$ for a (numerical valued) function which has the property that the individual

$$\left\{ \begin{array}{l} \text{will choose } f \text{ in preference to } g, \\ \text{be indifferent between } f \text{ and } g, \\ \text{or choose } g \text{ in preference to } f, \\ \text{according as } U(f) > U(g), \\ U(f) = U(g), \text{ or } U(f) < U(g). \end{array} \right. \quad (1)$$

Such a function will always exist, under mathematical assumptions of no practical consequence.

$U(f)$ thus gives rise to, or generates, an ordering of prospects and, if desired, may be called *the*, or in some respects better *a*, “utility function” of prospects. The individual may now be said to choose so as to maximize “utility.”

On this theory, the choice among the available alternatives depends only on their ordering. If there exists any one function U that is consistent with observed choices, then any other function that gives the same ordering of prospects will be equally consistent with observed choices. Consequently, it can equally well be said that the utility of prospects is given by any member of the family of functions

$$V[U(f)], \quad (2)$$

where V is an entirely arbitrary, strictly monotonic function, in particular

¹² *Op. cit.*, p. 61.

any function such that dV/dU is positive (so that all members of the family order the prospects in the same direction).

This is the central element of validity in Baumol's position. No one member of (2) has any more right to be called "the" utility function than any other member. In this sense, utility is not "measurable" no matter what may be the precise meaning of a prospect, or the class of prospects considered.

This "general" theory of choice is almost completely empty. It is not quite empty, because it requires consistency and transitivity, and very little more, and behavior is conceivable that would contradict these requirements; but it is so close to being empty that it is relatively useless for predicting behavior. A "special" theory consists in specifying more precisely the characteristics of $V[U(f)]$ or, equivalently, the ordering among the prospects. Let discussion be restricted to prospects that can be considered as probability combinations of elementary objects, and, for simplicity, consider the elementary objects to be amounts of income per unit time. Then a prospect is to be regarded as a set of alternative incomes, together with the associated probability that each income will be realized.¹³ The "special" theory under discussion is that there exists a function of income, say, $C(I)$, such that its expected value gives an ordering of prospects with property (1); that is, its expected value is *one* member of the family (2). Let us call this member U . Then, the special theory is that

$$U(f) = \bar{C}(I_f) \quad (3)$$

satisfies (1), where the bar over the C indicates expected value, and the subscript to the I indicates that the expected value is computed for the prospect f .¹⁴

If there exists any such function $C(I)$, then it is unique except for origin and unit of scale; that is, the only transformations of C that yield the same ordering of prospects consisting of more than one possible income with nonzero probability are of the form

$$D[C(I)] = s + tC(I), \quad (4)$$

with s arbitrary and $t > 0$.

This theory is by no means empty; indeed, if it were valid, knowledge of an individual's choices among some prospects each containing only one or two possible incomes would permit prediction of his choices among all other prospects, no matter how complex.

¹³ More precisely, let I be income per unit of time, and $P(I)$ the probability of receiving an income less than I . Then a prospect consists of a particular (cumulative) probability distribution, $P_f(I)$.

¹⁴ To state it more formally: $U(f) = U[P_f(I)] = \int_{-\infty}^{+\infty} C(I) dP_f(I)$.

If the theory is valid—that is, if it correctly predicts behavior—then there exists a function $C(I)$ unique except for origin and unit of scale. This function contains all the information that is relevant for predicting behavior. The way to use it is to calculate its expected value for alternatives under consideration and to predict that the alternative with the highest expected value will be chosen. It is conventional, but only conventional, to describe this process as the maximization of expected utility. If the hypothesis is valid, the expected value of $C(I)$ is one member of the family of functions defined by $V[U(f)]$. Admittedly, however, it is only one member. Any other member could be used: the cube, or fifth, power of the expected value of $C(I)$ will give the same ordering of the prospects as the expected value itself, and either could be termed “the” utility of prospects involving risk.

In our earlier article, which deals primarily with the consistency of this special theory with observed behavior, we, like others writing on this subject, called $D[C(I)]$, the family of functions given by (4), the “utility function of certain incomes.” If, however, one regards any member of (2) other than $\bar{C}(I)$ (or a linear transformation of it), say $V[I]$, as giving “the” utility of uncertain prospects, it is not valid to regard $C(I)$ as “the” utility of the certain income I ; “the” utility of the certain income I is then $V[C(I)]$. Our terminology was therefore misleading without more explanation than we gave and doubtless promoted confusion between the functions $D[C(I)]$ and the utility functions of prospects $V[U(f)]$. This confusion is manifest in Baumol’s note when he remarks: “If we . . . accept the view that any index obtained from a valid index by a monotone transformation is also valid, those of the Friedman-Savage results mentioned at the end of the preceding section, which refer to the shape of the income marginal utility curve, lose all their meaning.”¹⁵ This is a non sequitur: the results mentioned relate to the functions $D[C(I)]$; interpreted as referring to these functions, they retain all their meaning; they are essential for the content of the theory; but of course they are not directly relevant to any member of the family of functions $V[U(f)]$ other than the members given by $D[\bar{C}(I)]$.

What justification, then, is there for calling the particular members of $V[U(f)]$ given by $D[\bar{C}(I)]$, “the” utility of prospects involving risk and hence for speaking of utility as “measurable” rather than for calling the whole family of functions $V[U(f)]$ “the” utility of such prospects and hence speaking of utility as “ordinal” and not “measurable”? The justification is that, if the hypothesis is accepted, the former mode of speaking is far more convenient than the latter. Convenience may seem a slender justification; it is in fact an extremely important one. As Baumol correctly remarks, “In a sense any scale of measurement is arbitrary.

¹⁵ *Op. cit.*, p. 65.

Thus, aside from inconvenience, need anything be wrong with the use of a measure of distance which varies as the square . . . of the metric scale?"¹⁶ There is nothing wrong, aside from inconvenience; but inconvenience is not to be brushed aside so lightly. Aside from "inconvenience," need anything be wrong with the use of Roman rather than Arabic numerals; or with dropping numerical nomenclature entirely, replacing it by extemporaneous circumlocutions? Of course, the inconvenience of not using the metric scale cannot seriously be compared in magnitude with the inconvenience of not using mathematics in science at all, but this is only because measurements of length are just one among many kinds of measurements in science, not because of any difference in kind. And, even so, everyone in a country where the square of length as ordinarily measured was the official measure of "length" would suffer serious inconvenience indeed, for calculations like $(\sqrt{x} + \sqrt{y})^2$ and $(\sqrt{x} + \sqrt{y} + \sqrt{z})^2$ would have to be performed many times each day. All this computing of squares and square roots would be a pointless and heavy complication of empirical regularities that have been discovered in nature (only inadequately compensated by some convenience in dealing with right triangles).

If the expected-utility hypothesis is accepted, there is the same justification for calling "utility" "measurable," and $\bar{C}(I)$ its "measure," as there is for calling length and temperature "measurable," and the word "measurable" has precisely the same meaning in all three cases. Is it seriously being suggested by Baumol and those who share his views that the word now be redefined so as to apply to none of the three cases?

At the moment the "convenience" gained by adopting the convention of calling $D[\bar{C}(I)]$ "the" measure of utility is nothing like so clear or great as that involved in calling the ordinary sort of length "the" measure of length. This is partly because the hypothesis that makes the utility convention convenient is not firmly established, and partly because most discussion of the hypothesis is still at a highly abstract level at which there is no great inconvenience in using a broad rather than a narrow set of functions. The argument from convenience will be far stronger when, and if, repeated failures of the hypothesis to be contradicted strengthen confidence in its validity, and when, and if, it is sufficiently specialized to be useful in concrete applications. And if this hypothesis should be rejected because an alternative is found that is "better," in the sense of being equally fruitful and less frequently contradicted, then convenience may lead to the acceptance of a radically different "measure" of utility, or whatever new concept may replace it. The significant problem for positive economics is precisely to bring about such developments for this and alternative or supplementary hypotheses; to pro-

¹⁶ *Ibid.*

mote the construction of hypotheses about economic behavior that will enable predictions to be made about some behavior from the observation of other behavior. In this work, it will frequently be convenient to describe such a hypothesis in terms of some function with assigned properties and belonging to some class. "Measurability" is used to refer to the narrowness of the class. While this is only a convenient way of speaking, of describing a collection of rules for prediction, it is an important convenience, and economists should not be urged to forego its use simply because others may suppose that such statements refer to "reality" in an awesome spirit that is really nonsense. Economists can and should, like other scientists, recognize the relativity of their tools to the state of their science.

An excellent example for present purposes is provided by the attempts of Irving Fisher and Ragnar Frisch to measure utility referred to by Baumol.¹⁷ In its simplest form, for example, Fisher's method rests on a particular hypothesis about behavior, namely, that the ordering of various combinations of goods in the individual's scale of preferences can be reproduced by the sum of one-variable functions, each containing as a variable the quantity of a particular good. This hypothesis is by no means empty. If it were accepted, it would justify calling the sum of the one-variable functions (say, U) "the" utility function and would provide a "measurable" utility. Of course, it would nonetheless be true that any monotonic increasing function of U would give the same ordering of bundles of goods; however, only a linear one would be expressible as the sum of one-variable functions. But this particular hypothesis—of universal "independence" of goods in this sense—has as one implication that no good is an "inferior" good (i.e., that all income elasticities are positive), and this implication is contradicted by a great deal of evidence, which, indeed, is reasonably consistent with almost the opposite—that all goods, if narrowly defined, are "inferior" for some ranges of income. In consequence, the hypothesis of universal independence must be rejected. Less sweeping hypotheses requiring only limited "independence" can be formulated that would also provide a "measurable" utility. The difficulty is that an indefinitely large number can be formulated and that no one seems to have discovered a particular one that has fruitful empirical implications capable of being contradicted, yet not contradicted when submitted to test.

The failure of these experiments should be interpreted neither as a consequence of the nonmeasurability of utility in some absolute sense nor as showing that utility is not measurable. They were simply experiments that failed to produce empirical hypotheses acceptable to the authors themselves or to their colleagues in light of the available evidence.

¹⁷ *Ibid.*, p. 62.

It may be that future experiments along the same general lines will be more successful.

If the issue in question has seemed important, it is, we believe, largely because of a failure to distinguish sharply between “positive” economics and welfare economics. The discovery—if such it be—that a class of individual behavior can be predicted by supposing individuals to act as if they were maximizing the expected value of a function unique except for origin and unit of scale has, in and of itself, no welfare implication at all; and none is added by adopting the convention of calling the expected value of that function “utility.” As we remarked in our earlier paper, “it is entirely unnecessary to identify the quantity that individuals are to be interpreted as maximizing with a quantity that should be given special importance in public policy.”¹⁸ The ethical precept that society “should” promote the “welfare” of individuals is meaningless until “welfare” is given content. Any identification of “welfare” attained by individuals with “utility” as defined by the special theory of choice described above is itself an ethical precept, to be justified on ethical grounds, not a scientific proposition.

In light of the importance of utilitarianism as a philosophical creed and of its modern resurgence, particularly in “welfare economics,” it may well have promoted confusion to have used the term “utility” in a scientific theory of choice. But, even if this was an error, it hardly seems desirable to seek to correct it by denying to economists the use of the term “measurable” in the sense in which it is used in other sciences.

3. Conclusion

The hypothesis that individuals choose among alternatives involving risk as if they were seeking to maximize the expected value of some quantity, which has been called utility, is intended to be a scientific hypothesis enabling correct predictions to be made about individual behavior. It should be accepted—tentatively, of course, as all scientific hypotheses are—if it leads to “correct” predictions usually, or more frequently than any equally useful alternative; it should be rejected if its predictions are generally contradicted by observation. At the moment, the available evidence does not contradict the hypothesis, but it must be emphasized that the opportunities for contradiction have been few, so the direct evidence in its favor is meager. Confidence in the hypothesis derives largely from its coherence with the body of economic theory and, more important, from the plausibility of the postulates with which it can be shown to be equivalent rather than from repeated success in prediction.

If the hypothesis is accepted, it justifies or permits the definition of—

¹⁸ *Op. cit.*, p. 283, n. 11.

and therewith, the assignment of numerical values, unique except for origin and unit of scale, to—the quantity called “utility.” These values can in principle be determined by observing the choices made by an individual among a limited class of alternatives. The hypothesis, if accepted, thus makes it possible to regard “utility” as “measurable” in the same sense in which length and temperature are “measurable.” In all three cases this means the adoption of a convention on grounds of its convenience in using the hypotheses or theories in question. In no case does it justify regarding the particular “measure” as an unchanging “absolute” or as having any relevance for phenomena outside the range of those encompassed by the hypotheses on which the particular conventions rest. In all three cases it would be possible in principle to dispense with the convention and to adopt a more roundabout method of discourse, but only at the cost of a probably intolerable complication of known empirical regularities.

Two quite different classes of objections are made to this particular convention for assigning a “measure” to utility: (1) that the expected-utility hypothesis is not a useful or valid interpretation of actual behavior and (2) that any convention that would treat utility as measurable is undesirable or unnecessary. The first objection is, of course, unexceptionable; it can, and should, be removed only as and if experience in applying the hypothesis demonstrates that the hypothesis is fruitful and justifies acceptance. The second objection, on the other hand, cannot be accepted. It would deny economics a mode of speaking and expressing theories that has been found useful if not indispensable in other sciences. It expresses a viewpoint that its proponents would find impossible to apply consistently; they would have to deny themselves the privilege of speaking of money income, or stock of money, or size of population as measurable. If this viewpoint seems persuasive for utility, it is, we think, partly because some hypotheses that would, if verified, have implied convenient measures of utility were actually flagrantly contradicted by experience, and partly because of a widespread confusion between positive economics and normative economics, a widespread tendency to use the same word—utility—to stand for two quite different things: on the one hand, a quantity that it is useful to regard an individual as maximizing in interpreting his behavior and predicting his reactions to changed circumstances, and, on the other hand, a quantity that he “should” maximize or that society “should” maximize or help him maximize. The identification of these two conceptually different magnitudes is a maxim for wise behavior or an ethical precept, not a scientific proposition, and is in no way required by the adoption of a particular convention for “measuring” utility.

A Statistical Illusion in Judging Keynesian Models

with Gary S. Becker

Keynesian models have been used extensively in recent decades to predict and interpret movements in income. In its simplest form a Keynesian model contains one key relation, a consumption function connecting current expenditures on consumption with current income and perhaps other variables as well. This function yields an estimate of autonomous consumption and of the multiplier. These, in turn, yield an estimate of the level of income corresponding to any assumed or predicted level of investment (or, more generally, autonomous nonconsumption expenditures).

This model has been complicated in many ways; in all its variants, however, the consumption function plays a central role. In consequence, much attention has been devoted to empirical studies of the consumption function. The ultimate objective of predicting income has been replaced by the proximate objective of predicting consumption from current income and other variables. An unnoticed result of this shift in objective has been the adoption of a criterion for judging alternative consumption functions that, however relevant for the proximate objective, can be seriously misleading for the ultimate objective. The aim of this paper is to document this assertion and point out some of its implications. Since this can be done fully with the simplest model, and since the point in question has no direct relation to the controversy that has raged about the Keynesian approach, we shall assume throughout that the simple model described in the preceding paragraph is both theoretically and empirically applicable.

In judging alternative empirical consumption functions, the major criterion typically used is the error made on the average in estimating consumption from the function, albeit with some allowance for such

We are indebted to Phillip Cagan and Carl Christ for helpful comments.

considerations as the number of degrees of freedom used in estimating the function, the economic plausibility of the signs of the parameters, and so on. Suppose that two alternative empirical functions are estimated for the same historical period and are alike with respect to the subsidiary considerations and that the standard error of estimate of consumption is 5 per cent of the mean value of consumption for one function and 10 per cent for the other. Suppose that these functions are used to estimate retrospectively income in each year of the same historical period from the known magnitude of investment in that year. Does it follow that the first function will give more accurate estimates of income than the second? Surprisingly enough, the answer is "No." The error in the estimate of income depends not only on the accuracy of the estimate of consumption but also on the form of the consumption function—in particular, on the fraction of consumption which it designates as "autonomous."

Section I of this paper demonstrates this proposition and illustrates it with a number of empirical consumption functions. Section II gives a statistical interpretation of it, and Section III considers the effect of errors in estimating investment. Section IV shows that the choice of the form of the consumption function implies a specification of the dynamic characteristics of the economic structure, which is why it is of such basic importance.

I. The Relation between Errors in Predicting Consumption from Current Income and Errors in Predicting Income from Investment

Suppose that consumption is uniquely determined by income according to the relation

$$C = a + bY, \quad (1)$$

where C is aggregate consumption in a particular year, Y is aggregate income in the same year, and a and b are known positive constants. By definition

$$Y = C + I, \quad (2)$$

where I is realized investment (more generally, autonomous non-consumption expenditures) and by definition is also realized savings. Substituting equation (1) into equation (2) and solving for Y give

$$Y = \frac{a}{1-b} + \frac{I}{1-b}, \quad (3)$$

where $a/(1-b)$ is the amount of consumption that is autonomous of

investment,¹ and $1/(1 - b)$ is the multiplier applied to investment. The level of income can be predicted exactly from the level of investment. Similarly, the change in income resulting from a given change in government expenditures, treated as part of investment, can be predicted with certainty. This is the simple Keynesian model of our first paragraph in its most rigid form.

In application, allowance must be made for errors of measurement of the variables and for the effect of factors not included in the model. Perhaps the most common device is to treat equation (1) as a statistical relation rather than a functional relation, which is equivalent to replacing it by

$$C = a + bY + u, \quad (4)$$

where u is a random variable with a zero expected value. More sophisticated techniques allow in addition for systematic effects on consumption of other variables than current income; for example, equation (1) may be replaced by

$$C = a(Z) + bY + u, \quad (5)$$

where Z stands for a vector of other variables and $a(Z)$ for some unspecified function of these other variables.

According to equations (4) or (5), consumption is no longer uniquely determined by income or by income and the variables Z ; it is only determined up to an unexplained random component. To separate this analysis from problems of statistical estimation of parameters, assume for the moment that a , b , and the distribution of u are known without error. The expected value of C for a given value of Y , which is also the predicted value to be associated with this value of Y , is

$$C = a + bY, \quad (6)$$

where this, and the equations that follow, can be regarded as referring to both equations (4) and (5), provided only that a is interpreted in one case as a numerical constant and in the other as a number determined by Z .

The absolute error of prediction of C can be measured by the standard deviation of u ; the smaller this standard deviation, the more accurate the prediction. The relative error in predicting C can be measured by

¹ The term "autonomous consumption" is ambiguous, since it can refer either to the amount of consumption that is independent of investment or to the amount that is independent of income. In the latter sense, autonomous consumption is simply a , and the multiplier must be applied to it as well. We have tried to use the term "autonomous consumption" without further specification only when the statement in question holds for either meaning.

the standard deviation of the errors divided by the average value of C and is given by

$$V_C = \frac{\sigma_u}{\bar{C}} = \frac{\sigma_u}{a + b\bar{Y}}, \quad (7)$$

where σ_u is the standard deviation of u , and \bar{C} and \bar{Y} are the average values of C and Y , respectively. The quantity V_C measures the usefulness of equation (5) as a predictor of consumption.

From equations (2) and (4),

$$Y = \frac{a + I}{1 - b} + \frac{u}{1 - b}, \quad (8)$$

and

$$\hat{Y} = \frac{a + I}{1 - b}. \quad (9)$$

This predicted value of Y errs to the extent that $u/(1 - b)$ differs from zero. The relative error in predicting Y from a knowledge of I is

$$V_Y = \frac{\sigma_u}{(1 - b)\bar{Y}} = \frac{\sigma_u}{a + I}, \quad (10)$$

where \bar{I} is the average value of I . It follows from equations (7) and (10) that

$$V_Y = \frac{V_C(a + b\bar{Y})}{a + I} = \frac{V_C\bar{C}}{(1 - b)\bar{Y}}. \quad (11)$$

The relative error in predicting income is not determined solely by the relative error in predicting consumption or even by this together with the ratio of average consumption to average income. It depends also on the value of b . If $b = 0$,

$$V_Y = V_C \frac{\bar{C}}{\bar{Y}},$$

which, with positive savings, is less than V_C . As b rises and approaches 1, V_Y rises relative to V_C , equaling it when $b = 1 - \bar{C}/\bar{Y}$ and exceeding it for higher values of b .

Since in most empirical consumption studies, \bar{C}/\bar{Y} is of the order of 0.9, V_Y would exceed V_C when b exceeds about 0.1, which it generally does by a substantial margin. Given the level of average consumption and average income, a rise in b means a fall in a . Thus, the larger a , the smaller the error in income relative to the error in consumption.

Expressing consumption as the sum of a constant and a multiple of income involves distinguishing between consumption that is autonomous of income and consumption determined by current income. For a given level of average consumption and average income, the larger a , the larger the fraction of consumption that is autonomous.² Equation (11) therefore implies that, for a given error in predicting consumption, the error in predicting income by the method outlined is smaller, the more important is autonomous consumption, or, what is equivalent, the weaker is the connection between current income and current consumption.

Why should the error in predicting income differ from and usually be larger than the error in predicting consumption? The answer is that, in the model under consideration, a given error in predicting consumption is magnified by the multiplier process into a much larger error in predicting income. Hence, the larger the multiplier, the larger the resulting error in predicting income.

Several consumption functions have been fitted by one of us to annual time-series data on income and consumption for the United States for 1905–51 (excluding war years). These functions differ considerably in the division they imply between autonomous and induced consumption and hence will serve to illustrate the importance of this division. Let $C(t)$ stand for real personal consumption per capita at time t , and $Y(t)$ for real disposable income per capita at time t . Four functions fitted are:

$$C(t) = a_1 + b_1 Y(t) + u_1, \quad (12)$$

$$C(t) = b_2 Y(t) + u_2, \quad (13)$$

$$C(t) = k\beta \int_{-\infty}^t e^{(\beta-a)(T-t)} Y(T) dT + u_3, \quad (14)$$

and

$$C(t) = a_2 + A e^{at} + u_4, \quad (15)$$

where the u 's are random disturbances, and a_1 , a_2 , b_1 , b_2 , A , k , β , and α are parameters estimated from the data by least squares. The numerical values of these parameters are given in Table 1. In fitting function (14), the integral was approximated by a summation of a limited

² Since $a/(1-b)$ increases with a , so long as b is less than unity, which it almost invariably is, the statement in the text is valid for either of the interpretations of autonomous consumption noted in n. 2.

number of terms, each for a calendar year.³ The idea behind function (14) is that consumption depends not on current income alone but on a longer-term concept of income, "permanent" or "expected" income. And expected income in turn is approximated by a weighted average of all past incomes, adjusted for secular trend (this is the role of a), the weights declining exponentially as one goes back in time.

The observed relative error in estimating consumption by each of these functions is given in column (1) of Table 1. Function (14) on the average fits the consumption data best. It is, however, only moderately better than functions (12), (13), and (15), which follow in that order.

In order to provide a standard for judging these relative errors, two functions involving "naïve" predictions of consumption are also covered by the table:

$$C(t) = C(t-1) + u_5, \quad (16)$$

and

$$C(t) = e^a C(t-1) + u_6. \quad (17)$$

These functions yield roughly the same relative error as function (15), and are about equal to it in their predictive accuracy. Function (15) estimates consumption from its own trend as approximated by an exponential.

These consumption functions can be used to predict income from investment by the procedure outlined earlier: the consumption function is substituted into the income identity, and the resulting equation is solved for current income as a function of current investment and of any other variables that enter into the consumption function. Column (2) of Table 1 gives the relative errors in predicting income in this way.

These numerical results strikingly illustrate the importance of the division of consumption into autonomous and induced parts. Function (13) implies that all consumption is induced; in the previous notation, that $a = 0$. This means that errors in consumption are multiplied manyfold when autonomous expenditures are used to predict income; the relative error is roughly seven times as large in predicting income as in predicting consumption during this period. Functions (15), (16), and (17) imply that all consumption is autonomous, that $b = 0$. This means that there is no multiplication of errors in consumption in predicting income. The relative error must be less on the average in predicting income than in predicting consumption. During the period 1905–51 it

³ See Milton Friedman, *A Theory of the Consumption Function*, a publication of the National Bureau of Economic Research (Princeton, N.J.: Princeton University Press, 1957). The basic data are from Raymond Goldsmith, *A Study of Savings* (Princeton, N.J.: Princeton University Press, 1955), Vols. I, II, and III.

TABLE 1
 ERRORS IN THE PREDICTION OF CONSUMPTION AND INCOME FROM DIFFERENT CONSUMPTION FUNCTIONS

CONSUMPTION FUNCTION	RELATIVE ERROR IN PREDICTING			
	Consumption from a Known Level of Income (1)	Income from a Known Level of Investment (2)	Income from an Estimated Level of Investment* (Consumption Known with Certainty) (3)	Income from Estimated Consumption and Investment* (4)
(12) $C(t) = 70.2 + .780Y(t) + u_1$.049	.196 [†]	.100	.242 [‡]
(13) $C(t) = .890Y(t) + u_2$.057	.462 [†]	.200	.551 [‡]
(14) $C(t) = (.879) (.4) \int_{-\infty}^{t-\infty} e^{.38(T-t)} Y(T) dT + u_3$.040	.051	.029	.059
(15) $C(t) = 162.0 + 237.0 e^{.02t} + u_4$.066	.059	.020	.062
(16) $C(t) = C(t-1) + u_5$.068	.061	.020	.065
(17) $C(t) = e^{.02t} C(t-1) + u_6$.064	.057	.020	.061

* The relative error in predicting investment is assumed to be .20.
[†] Instead of .196 and .462, these figures would be .147 and .430, respectively, if more efficient estimates of the parameters of these consumption functions had been used.
[‡] Instead of .242 and .551, these figures would be .178 and .465, respectively, if more efficient estimates of the parameters of these consumption functions had been used.

averaged about nine-tenths as large. Functions (12) and (14) assume that some consumption is autonomous and some induced; neither a nor b is zero. The numerical values assigned to the parameters of function (12) for this period imply that roughly one-eighth of consumption is autonomous of income and seven-eighths is induced, so there is substantial multiplication of consumption errors. The average percentage error is three times as large for the income predictions as for the consumption predictions. For function (14), on the other hand, the numerical values assigned to the parameters imply that about two-thirds of consumption is autonomous of income and that less than one-third is induced. The multiplier is small, so that the relative error is only about 25 per cent larger for the income predictions than for the consumption predictions.

Despite the apparently good showing of functions (12) and (13) in predicting consumption, they are worthless in predicting income. The simplest "naïve" model does very much better.

II. A Statistical Interpretation of the Results

The errors in column (2) of Table 1 are for predictions obtained by substituting the estimated consumption functions into the income identity. While this is the procedure suggested by the simple non-stochastic model in which consumption is functionally related to income—the model summarized in equations (1), (2), and (3)—it is not the optimum procedure when random disturbances are introduced. This can be illustrated most conveniently with functions (12) and (13). From the income identity, $C = Y - I$. Substitute this expression for C in functions (12) and (13) and solve for I . The result is

$$I = a_1 + (1 - b_1)Y - u_1 = a' + b'_1Y + u_1, \quad (18)$$

for function (12), and

$$I = (1 - b_2)Y - u_2 = b'_2Y + u_2 \quad (19)$$

for function (13).

Since the independent variable is the same as in functions (12) and (13) and the disturbances are changed only in sign, the values of the parameters that minimize the sum of the squared disturbances in (12) and (13) will also minimize the sum of the squared disturbances in (18) and (19). Estimating the parameters in (12) and (13) by least squares is therefore equivalent to estimating the parameters in (18) and (19). But (18) and (19) are obviously the wrong regressions to use in estimating Y from a value of I that is assumed to be known without error.

For this purpose, one should use the regression of Y on I , not of I on Y . That is, one should use

$$Y = d + e_1 I + w_1, \quad (20)$$

or

$$Y = e_2 I + w_2. \quad (21)$$

The standard errors of estimate of Y using functions (20) and (21) are necessarily smaller than those of Y using functions (12) and (13), the relation between the corresponding relative errors being given by⁴

$$V_Y^{(20)} = r_{IY} V_Y^{(12)}, \quad (22)$$

and

$$V_Y^{(21)} = r'_{IY} V_Y^{(13)}, \quad (23)$$

where r_{IY} is the usual correlation coefficient between I and Y , and r'_{IY} is the correlation coefficient obtained by treating I and Y as if they had means of zero, that is

$$r'_{IY} = \frac{\sum IY}{(\sum Y^2 \sum I^2)^{1/2}}. \quad (24)$$

For the period 1905–51, $r = .75$ and $r' = .91$. It follows that the use

⁴ *Proof:* Using function (12) implies predicting Y from

$$\hat{y} = \frac{1}{b_1} i, \quad (i)$$

where y and i stand for the deviations of y and I from their mean values and where

$$b_1 = \frac{\sum iy}{\sum y^2}. \quad (ii)$$

The sum of the squares of the deviations of the observed values of Y from the estimates given by equation (i) is

$$\begin{aligned} \sum \left(y - \frac{1}{b_1} i \right)^2 &= \sum y^2 - \frac{2}{b_1} \sum iy + \frac{1}{b_1^2} \sum i^2 \\ &= \sum y^2 - 2 \sum y^2 + \frac{1}{r^2} \sum y^2 \\ &= \frac{(1 - r^2)}{r^2} \sum y^2, \end{aligned} \quad (iii)$$

where r stands for r_{IY} , or the correlation between income and investment. By the usual formula, the corresponding sum of squared deviations in predicting Y from equation (20) is

$$\sum (Y - e_1 I)^2 = (1 - r^2) \sum Y^2. \quad (iv)$$

Comparison of equations (iii) and (iv) gives equation (22). Equation (23) can be proved in the same way.

of the correct regression would have yielded a relative error for the analogue to function (12) of .147 instead of .196, and for the analogue to function (13) of .420 instead of .462. Similarly, use of the correct regression would have reduced the error for function (14) below .051, though it is somewhat more difficult to estimate by how much, and we have not done so. The remaining functions are not affected by this statistical defect, so the errors for them need no modification. The use of the wrong regression clearly accounts for only a small part of the initial difference among the various functions in the prediction of income. Functions (12) and (13) remain by all odds the poorest predictors of income, although they appear to be the second and third best predictors of consumption.

While these statistical considerations do not resolve this apparent conflict, they show, in a rather different way from that we have shown previously, how it arises. The relatively small errors made by functions (12) and (13) in predicting consumption from income are in part a statistical illusion.

In these functions, current consumption is related to current income, of which it is itself the major part. Consumption is, as it were, being correlated with itself. The fact that knowledge of *income* enables us to estimate consumption fairly accurately in no way implies that knowledge of that part of income equal to *investment* also enables us to estimate consumption accurately. To determine whether this is so, we must relate consumption not to income but to investment alone, as is implicitly done in equations (20) and (21). These yield precisely the estimates of income that would be obtained if consumption were estimated from the least-squares regression of consumption on investment and then the (assumed) known investment were added to the resulting estimate of consumption.

Given a positive correlation between investment and consumption, the error in estimating consumption from investment is bound to be larger than the error in estimating consumption from income.⁵ Indeed, the relative error is bound to be larger than the relative error in estimating income from investment as in equations (20) and (21), the reason being that these equations incorporate the same kind of a statistical illusion as functions (12) and (13): income is there correlated with part of itself. Put differently, a prediction of income by the kind

⁵ If $r_{CI} = \pm 1$, then $r_{CY} = \pm 1$; otherwise, regardless of the sign of the correlation between investment and consumption, $r_{CY} > r_{CI}$. However, it is not necessarily true that $r_{CY} > r_{CI}$. If r_{CI} is negative and sufficiently large in absolute value (greater in absolute value than $\frac{1}{2}[\sigma_C/\sigma_I]$), then $r_{CY}^2 < r_{CI}^2$, which is to say that a prediction of consumption from investment would be more accurate than one from income, because the negative correlation between investment and consumption would blur sufficiently the perfect positive correlation between consumption and itself.

of model we have been using throughout consists essentially of two steps: first, prediction of investment; second, prediction of consumption from investment. Since we have been assuming that the first step is without error, the only source of error from the model is the second. The relative error in this step is .22 for a function like (12) and .53 for a function like (13), with I substituted for Y . It now no longer seems so surprising that these functions should do so much worse than the others in predicting income.

Function (14), like functions (12) and (13), is also infected by a statistical illusion in predicting consumption. It too partly relates consumption to a total including itself. But the illusion is in this case very much smaller. The reason of course is the one already pointed out. Current income plays a much smaller role in estimating consumption in function (14) than in either function (12) or function (13). The marginal propensity to consume out of *current* income is .78 in function (12), .89 in function (13), but only .29 in function (14). In consequence, if investment were substituted for current income in computing function (14), the relative error would be raised to something less than .058.⁶

The remaining equations are free from any illusion, for current income does not enter into them at all. The absence of an illusion is reflected in errors that are lower in column (2) than in column (1), the relation one would expect to hold when income is derived from consumption by adding a value for investment that is assumed to be known without error.

It is tempting at first sight to attribute the great superiority of function (14) over function (12) in predicting income to the inclusion of additional variables, namely, incomes in earlier years.⁷ It is clear, however, from functions (15), (16), and (17) that this is only a minor source of its better performance. These functions have the same or a smaller number of variables and the same or a smaller number of parameters estimated from the data. Yet they do almost as well as function (14).

III. Errors in Predicting Investment

So far we have been assuming investment (or, more generally, autonomous non-consumption expenditures) to be known without error. Let

⁶ This is an approximation derived by multiplying .051 by the ratio of average income to average consumption, since we have not computed what the error would be if the correct regression, which would give an error less than 0.51 in column (2), were used.

⁷ Function (14) looks as if it had one more parameter than function (12). In fact, it has only two independent parameters. The parameter a (set equal to .02) is not "identified"; it is inserted from external evidence to give to other parameters a more meaningful interpretation, but its use does not improve the fit.

us drop this assumption and suppose that investment is estimated from

$$I(t) = g\{x_1(t), x_2(t), \dots, x_n(t)\} + w, \quad (25)$$

where $x_1(t), \dots, x_n(t)$ stand for the values of variables other than income and consumption at time t , and w is the error in predicting I from g . The relative error in predicting I is equal to

$$V_I = \frac{\sigma_w}{\bar{g}}. \quad (26)$$

If we combine equation (25) with equations (2) and (4) and solve for Y , the result is

$$Y = \frac{a + g}{1 - b} + \frac{u + w}{1 - b}. \quad (27)$$

If we assume that u is uncorrelated with w , we get

$$V_Y^2 = \frac{\sigma_u^2 + \sigma_w^2}{(1 - b)^2 \bar{Y}^2} = \frac{V_C^2 \bar{C}^2}{(1 - b)^2 \bar{Y}^2} + \frac{V_I^2 \bar{I}^2}{(1 - b)^2 \bar{Y}^2} \quad (28)$$

or

$$V_Y = \left\{ \frac{V_C^2 \bar{C}^2}{(1 - b)^2 \bar{Y}^2} + \frac{V_I^2 \bar{I}^2}{(1 - b)^2 \bar{Y}^2} \right\}^{1/2}. \quad (29)$$

From the symmetrical way in which a and g and u and w enter into equation (27), it is clear that errors in investment play the same role as errors in consumption and that our preceding analysis for the special case in which consumption alone is subject to error carries over completely to the special case in which investment alone is subject to error. It is only necessary to replace a by g , u by w , and \bar{C} by \bar{I} . Some illustrative calculations for this special case, parallel to those already presented, are given in column (3) of Table 1. These calculations assume throughout a relative error in investment of 20 per cent ($V_I = .20$)—this figure is entirely arbitrary, since we have no empirical estimate of the errors in predicting investment comparable to the estimates for consumption in column (1)—and a ratio of mean investment to mean income of 0.1, roughly the observed ratio for the period in question.

Functions (12) and (13) again imply much larger errors in predicting income than the other functions: for function (13) the percentage error in income is the same as in investment; for function (12) it is half as large. The implied errors are drastically smaller for functions (14) through (17), which do not differ markedly among themselves. The 20 per cent error assumed for investment implies only a 2 or 3 per cent error in predicting income, a result that is intuitively plausible. Investment is a small fraction of total income, approximately one-tenth for

this period; in consequence, a given error in investment is a much larger percentage of investment than of income. In functions (12) and (13), this effect is countered by the multiplier process, which multiplies the error as well as investment proper. In the remaining functions this multiplication of the error is largely or entirely absent, since these functions involve a small or zero multiplier.

The error in predicting income when both consumption and investment are subject to error, given by equation (29), is equal to the square root of the sum of the squares of two partial errors in predicting income, one being the error in predicting income from known investment with consumption a stochastic function of income and specified other variables, the other being the error in predicting income from estimated investment with consumption a non-stochastic function of income and specified other variables. The determinants of these partial errors have already been discussed: the numerical magnitude of V_I is determined by the investment function, of V_C and of b by the consumption function, and of \bar{C} relative to \bar{Y} by the rate at which a society is adding to its wealth. For the United States during this period, consumption averaged about nine-tenths of income. This means that the error in predicting income caused by V_C would equal the error caused by V_I only if V_I equaled $9V_C$.

The analysis can again be illustrated by functions (12) through (17). As before, we arbitrarily suppose V_I to be equal to .20. The value of V_C is given by column (1) of Table 1. Column (4) gives the resulting values of V_Y for each consumption function. Functions (12) and (13) again yield by far the worst predictions of income.⁸ Function (14) yields slightly better predictions than functions (15) through (17). The "naïve" models of consumption yield decidedly better predictions of income, on the average, than some more sophisticated models and only slightly worse predictions than others.⁹

⁸ The relative errors in column (4) for functions (12) and (13) are based on inefficient estimates of the parameters in these consumption functions. If efficient estimates were used, the relative errors would be obtained from the residual errors of the regression of Y on g . Sample values of g are not available; therefore, we could not compute these errors. The relative errors for functions (12) and (13) in n. † to Table 1 were obtained by taking the square root of the sum of the squares of the relative error in predicting Y from I (n. †) and the relative error in predicting Y caused by the assumed error (.20) in predicting I from g (col. [3]).

⁹ Completely "naïve" predictions of income would involve models like

$$Y_t = Y_{t-1}, \quad (\text{v})$$

or

$$Y_t = e^{.02} Y_{t-1}. \quad (\text{vi})$$

The relative errors, in predicting income for these models are

$$V_Y^{(\text{v})} = .077 \quad \text{and} \quad V_Y^{(\text{vi})} = .075.$$

As noted earlier, the prediction of income by means of a Keynesian model of the kind we have been discussing can be viewed as consisting of two related steps: first, the prediction of investment; second, the prediction of consumption from investment and other variables. Since investment is small relative to income, the final error in predicting income is dominated by the error made in the second step. Even a 30 per cent error in predicting investment means at most a 3 or 4 per cent error in income via the first step. And the error in the second step depends not only on the error in investment but also on the kind of consumption function employed. In consequence, the major requisite to the improvement of prediction by the use of such models may not be, as is so often supposed, an improvement in the accuracy with which investment is predicted. It may well be more important to improve the consumption function used to predict consumption from investment. And it hardly needs repetition at this point than the relevant criterion of improvement is not the error made in predicting consumption from *income* and other variables but in predicting consumption from *investment* and other variables.

IV. An Economic Interpretation of the Results

The basic difference among the models whose predictive value we have been examining is not in their statistical characteristics—the number of variables or parameters or the mathematical form of the function—but in the structure they attribute to the economic system. At the one extreme, function (13), which takes consumption to be proportional to income except for random shocks, combined with the acceptance of investment as *the* autonomous variable, assumes that the only source of continuity in the system is any continuity in investment itself. A change in investment propagates itself through the rest of the system with a sizable multiplier; and, in the original form of function (13), the same multiplier applies to random disturbances in consumption. Income is at best as stable, or unstable, as investment and, given any consumption disturbances, more unstable.

At the other extreme is function (15), which takes consumption to be determined primarily by its own trend except for random disturbances. Given that investment is autonomous, neither its changes nor random disturbances in consumption are multiplied. Instead, the two are added together and so tend to average out; income is simply the sum of two independent components. The large trend element in consumption introduces a stable background around which these perturbations occur. The instability of income is on the average less than the average instability of investment and consumption taken separately.

The other functions are between these conceptual extremes. Function (12) introduces an element of continuity that is lacking from function (13), namely, autonomous consumption; even though this is a relatively small part of total consumption, it is enough to give considerably greater stability to the system. Function (14) introduces, in addition, incomes in earlier years as a determinant of consumption and thereby supposes a persistence over time that is entirely lacking from function (13) and largely lacking from function (12). The persistence is not complete: current investment plays a role; random disturbances in consumption and changes in investment are multiplied but only mildly so.

These statements are qualitative and not rigorous, but perhaps they suffice to show that the choice of a form for a consumption function in a Keynesian model implicitly involves a specification of the dynamic characteristics of the economic system itself. And perhaps they also suffice to show that calculations like those in column (2) of Table 1 can have important substantive and not merely methodological interest.

From this viewpoint, the most striking result, it seems to us, is that function (15), implying the greatest persistence in consumption and complete insulation of consumption from changes in investment, does nearly as well in reproducing the actual movements of income as any of the other functions and dramatically better than the two functions, (12) and (13), that give the most scope to the investment-multiplier process. Regarding investment as having no multiplier effects is a much better first approximation than regarding investment as the prime mover, at least with respect to the changes in real per capita income from 1905 to 1951 as recorded in the data from which the functions in Table 1 were computed.

This conclusion cannot be regarded as firmly established. The functions used in Table 1 were chosen because they were handy, having been computed for another purpose. We initially intended them to illustrate solely the methodological issue involved and were ourselves surprised at how striking a substantive conclusion they suggest. The concepts employed may not be the best for the substantive purpose: in particular, it would be desirable to use a broader concept of investment or autonomous non-consumption expenditures—these data correspond to personal realized savings and so exclude, for example, investment financed by undistributed corporate earnings. It might be desirable to allow for the effects of built-in stabilizers. It might also be desirable to allow for price movements in some way other than by using deflated income and consumption. The relative error might be replaced or supplemented by other measures of the error in prediction. The period employed and the time unit may not be the best. It might be preferable to use quarterly or monthly data for shorter periods; it might be desirable to introduce time lags between consumption and investment; and there are doubtless

still other possible modifications that will occur to the reader. At the same time, some experiments made for another purpose lead us to doubt that the major conclusion would be much affected, at least for the United States.¹⁰

Whether our substantive conclusion about the structure of the economy is confirmed or not, the results in Table 1 suffice to establish firmly a substantive conclusion about economists. The propensity of economists to judge Keynesian models by measures like those in column (1) of Table 1 has led us to accept a specification of the character of the economic structure that our data do not support. The finding that consumption can be predicted statistically from current income alone with a relatively small error reinforced the view that the corresponding simple Keynesian models incorporated a basically correct specification of the economic structure, although one that was unduly simple. If our analysis is valid, this widespread belief that consumption and income can be regarded as dancing primarily to the tune of unstable investment is largely the product of a statistical illusion.

V. Summary

Our major conclusions can be set forth very briefly.

1. The accuracy with which a consumption function enables consumption to be predicted from current *income* is a poor criterion of the usefulness of this consumption function in predicting income from investment via a Keynesian model. In general, for a given accuracy in predicting consumption from income, a consumption function gives better results in predicting income, the lower the implicit investment multiplier, that is, the larger the fraction of consumption regarded as autonomous of investment.

2. The relevant criterion is the accuracy with which consumption can be predicted from current *investment*, not from income, and the accuracy of the two predictions may not be at all closely related.

3. Some illustrative calculations suggest that errors in estimating investment tend to have less of an effect on the accuracy of predictions of income than does the use of an inappropriate consumption function. Improvement in the consumption function, especially in the prediction of consumption by variables other than investment, may therefore make a greater contribution to our ability to predict income than improvement in the estimates of investment.

4. The choice of a consumption function implies a specification of the dynamic characteristics of the economic structure. Illustrative cal-

¹⁰ The experiments have been made in connection with studies under way in the Workshop on Money and Banking of the University of Chicago.

culations for the United States for 1905–51 suggest that it is a better first approximation to regard real investment as having no multiplier effects on real consumption whatever and real consumption as simply determined by its own long-term trend, than to regard real investment as the prime mover and changes in real consumption and hence in real income as largely mirroring changes in real investment. The widespread belief to the contrary—in so far as it derives from statistical evidence on the consumption function—is largely the product of a statistical illusion.

The Demand for Money: Some Theoretical and Empirical Results¹

In countries experiencing a secular rise in real income per capita, the stock of money generally rises over long periods at a decidedly higher rate than does money income. Income velocity—the ratio of money income to the stock of money—therefore declines secularly as real income rises. During cycles, to judge from the United States, the only country for which a detailed analysis has been made, the stock of money generally rises during expansions at a lower rate than money income and either continues to rise during contractions or falls at a decidedly lower rate than money income. Income velocity therefore rises during cyclical expansions as real income rises and falls during cyclical contractions as real income falls—precisely the reverse of the secular relation between income and velocity.

These key facts about the secular and cyclical behavior of income velocity have been documented in a number of studies.² For the United States, Anna Schwartz and I have been able to document them more fully than has hitherto been possible, thanks to a new series on the stock of money that we have constructed, which gives estimates at annual or semiannual dates from 1867 to 1907 and monthly thereafter. This fuller documentation does not, however, dispel the apparent contradiction

¹ This paper reports on part of a broader study being conducted at the National Bureau of Economic Research by Anna J. Schwartz and myself. I am indebted to Mrs. Schwartz for extensive assistance and numerous suggestions in connection with the present paper.

This paper has been approved for publication as a report of the National Bureau of Economic Research by the Director of Research and the Board of Directors of the National Bureau, in accordance with the resolution of the board governing National Bureau reports (see the *Annual Report of the National Bureau of Economic Research*). It is to be reprinted in the National Bureau's series of "Occasional Papers."

² See in particular Richard T. Selden, "Monetary Velocity in the United States," in Milton Friedman (ed.), *Studies in the Quantity Theory of Money* (Chicago: University of Chicago Press, 1956), pp. 179–257; and Ernest Doblin, "The Ratio of Income to Money Supply: An International Survey," *Review of Economics and Statistics*, August, 1951, p. 201.

between the secular and the cyclical behavior of income velocity. On the contrary, as the summary of our findings in the following section makes explicit, it reveals an additional contradiction or, rather, another aspect of the central contradiction.

Previous attempts to reconcile the secular and cyclical behavior of the velocity of circulation of money have concentrated on variables other than income, such as the rate of interest or the rate of change of prices. These attempts have been unsuccessful. While such other variables doubtless affect the quantity of money demanded and hence the velocity of circulation of money, most do not have a cyclical pattern that could explain the observed discrepancy. In any event, it seems dubious that their influence on velocity is sufficiently great to explain so large a discrepancy.

An alternative theoretical explanation of the discrepancy is suggested by the work I have done on consumption—a rather striking example of how work in one field can have important implications for work in another that has generally been regarded as only rather distantly related. This theoretical explanation, which concentrates on the meaning attached to “income” and to “prices,” is presented in Sections II and III below and turns out to be susceptible of quantitative test. The quantitative evidence in Section IV is highly favorable. The result is both a fuller understanding of the observed behavior of velocity and a different emphasis in the theory of the demand for money.

One important feature of monetary behavior not accounted for by this explanation is the consistent tendency for actual cash balances, adjusted for trend, to lead at both peaks and troughs in general business. In Section V, a preliminary attempt is made to explore factors that might account for the discrepancy between desired cash balances as determined by income alone and actual cash balances. Finally, in Section VI, some broader implications of the results presented in this paper are explored.

I. A Summary of the Empirical Evidence for the United States

A full documentation of our findings about the secular and cyclical behavior of the stock of money and its relation to income and prices will be given in a nearly completed National Bureau of Economic Research monograph by Anna J. Schwartz and myself. For present purposes, a brief summary of a few of our findings will suffice.

A. Secular Behavior

1. Secular changes in the real stock of money per capita are highly correlated with secular changes in real income per capita. In order to

study this relation, we have used average values over complete reference cycles as our elementary observations. For twenty cycles measured from trough to trough and covering the period from 1870 to 1954, the simple correlation between the logarithm of the real stock of money per capita and the logarithm of real income per capita is 0.99, and the computed elasticity is 1.8.³

A 1 per cent increase in real income per capita has therefore, on the average, been associated with a 1.8 per cent increase in real cash balances per capita and hence with a 0.8 per cent *decrease* in income velocity. If we interpret these results as reflecting movements along a stable demand relation, they imply that money is a "luxury" in the terminology of consumption theory. Because of the strong trend element in the two series correlated, the high correlation alone does not justify much confidence that the statistical regression is a valid estimate of a demand relation rather than the result of an accidental difference in trends. However, additional evidence from other sources leads us to believe that it can be so regarded.

We have investigated the influence of both rates of interest and rates of change of prices. In our experiments, the rate of interest had an effect in the direction to be expected from theoretical considerations but too small to be statistically significant. We have not as yet been able to isolate by correlation techniques any effect of the rate of change of prices, though a historical analysis persuades us that such an effect is present.

2. Over the nine decades that we have studied, there have been a number of long swings in money income. As a matter of arithmetic, these swings in money income can be attributed to movements in the nominal stock of money and in velocity. If this is done, it turns out that the swings in the stock of money are in the opposite direction from those in velocity and so much larger in amplitude that they dominate the movements in money income. As a result, the long swings in prices mirror faithfully the long swings in the stock of money per unit of output. These long swings are much more marked in money income and in the nominal stock of money than in real income and in the real stock of money, which is to say that the long swings are largely price swings.

³ The corresponding figures for cycles measured from peak to peak are 0.99 and 1.7. In these and later correlations, "money" is defined as including currency held by the public, adjusted demand deposits, and time deposits in commercial banks. This total is available for the period from 1867 on, whereas the total exclusive of time deposits is not available until 1914. For other reasons supporting our definition see the NBER monograph now in preparation. For income, we have used Simon Kuznets' estimates of net national product adjusted for wartime periods to a concept approximating that underlying the current Department of Commerce estimates, and for prices, the deflator implicit in Kuznets' estimates of net national product in constant prices.

B. *Cyclical Behavior*

1. The real stock of money, like real income, conforms positively to the cycle; that is, it tends to rise during expansions and to fall, or to rise at a less rapid rate, during contractions. However, the amplitude of the movement in the real stock of money is decidedly smaller than in real income. If we allow for secular trends, a 1 per cent change in real income during a cycle is accompanied by a change in the real stock of money in the same direction of about one-fifth of 1 per cent.

It follows that income velocity tends to rise during cyclical expansions when real income is rising and to fall during cyclical contractions when real income is falling—that is, to conform positively. So far as we can tell from data that are mostly annual, velocity reaches both its peak and its trough at roughly the same time as general economic activity does.

2. Cyclical movements in money income, like the long swings, can be attributed to movements in the nominal stock of money and in velocity. If this is done, it turns out that the movements in the stock of money and in velocity are in the same direction and of roughly equal magnitude, so that neither can be said to dominate the movements in money income.

3. Table 1 summarizes the size of the cyclical movements in the variables used in the analysis, where the size of cyclical movement is measured by the excess of the rate of change per month during cyclical expansions over that during cyclical contractions.

C. *The Contrast*

These findings are clearly in sharp contrast. Over long periods, *real* income and velocity tend to move in opposite directions; over reference cycles, in the same direction. Over long periods, changes in the nominal stock of money dominate, at least in a statistical sense, the swings in *money* income, and the inverse movements in velocity are of minor quantitative importance; over reference cycles, changes in velocity are in the same direction as changes in the nominal stock of money and are comparable in quantitative importance in accounting for changes in money income. I turn to an attempted reconciliation.

II. A Suggested Explanation

It is important to note at the outset an essential difference between the determinants of the nominal stock of money, on the one hand, and the real stock of money, on the other. The nominal stock of money is determined in the first instance by the monetary authorities or institutions

TABLE 1
CYCLICAL MOVEMENTS IN INCOME, MONEY STOCK, INCOME VELOCITY, AND PRICES:
DIFFERENCE IN MONTHLY RATE OF CHANGE BETWEEN REFERENCE EXPANSION
AND CONTRACTION, ANNUAL ANALYSIS, 1870–1954, EXCLUDING WAR CYCLES*

	CHANGE PER MONTH IN REFERENCE-CYCLE RELATIVES DURING REFERENCE		EXCESS OF EXPANSION OVER CONTRACTION
	Expansion	Contraction	
	(1)	(2)	
Twelve mild depression cycles:			
Money income	0.64	−0.07	0.71
Money stock	.55	.28	0.27
Income velocity	.08	−.32	0.40
Implicit price deflator	.12	−.02	0.14
Real income	.52	−.05	0.57
Real stock of money	.43	.30	0.13
Six deep depression cycles:			
Money income	.64	−.97	1.61
Money stock	.60	−.28	0.88
Income velocity	.02	−.69	0.71
Implicit price deflator	.16	−.44	0.60
Real income	.46	−.53	0.99
Real stock of money	0.42	0.18	0.24

* The series were analyzed as described in A. F. Burns and W. C. Mitchell, *Measuring Business Cycles* (New York: National Bureau of Economic Research, 1947), pp. 197–202. Because of rounding, col. 3 sometimes disagrees with the difference between cols. 1 and 2. Deep depression cycles are 1870–78, 1891–94, 1904–8, 1919–21, 1927–32, and 1932–38. All others are mild depression cycles except for war cycles 1914–19 and 1938–46, which are excluded. The basis of classification is described in the NBER monograph on the money supply now in preparation. *Money income* is net national product at current prices, preliminary estimates by Simon Kuznets, prepared for use in the NBER study of long-term trends in capital formation and financing in the United States. Variant III (from 1929 based on estimates of commodity now and services prepared by the Department of Commerce). *Money stock* is averaged to center on June 30 from data in the money monograph just mentioned. *Income velocity* is money income divided annually by money stock. *Implicit price deflator* is money income divided by real income. *Real income* is net national product, 1929 prices. Variant III from the same source as money income. *Real stock of money* is money stock divided by the implicit price deflator.

and cannot be altered by the non-bank holders of money. The real stock of money is determined in the first instance by the holders of money.

This distinction is sharpest and least ambiguous in a hypothetical society in which money consists exclusively of a purely fiduciary currency issued by a single money-creating authority at its discretion. The nominal number of units of money is then whatever amount this authority creates. Holders of money cannot alter this amount directly. But they can make the real amount of money anything that in the aggregate they want to. If they want to hold a relatively small real quantity of money, they will individually seek to reduce their nominal cash balances by increasing expenditures. This will not alter the nominal stock of money to be held—if some individuals succeed in reducing their nominal cash balances, it will only be by transferring them to others. But it will raise the flow of expenditures and hence money income and prices and thereby reduce the real quantity of money to the desired level. Conversely, if they want to hold a relatively large real quantity of money, they will individually seek to increase their nominal cash balances. They

cannot, in the aggregate, succeed in doing so. However, in the attempt, they will lower the nominal flow of expenditures, and hence money income and prices, and so raise the real quantity of money. Given the level of real income, the ratio of income to the stock of money, or income velocity, is uniquely determined by the real stock of money. Consequently, these comments apply also to income velocity. It, too, is determined by the holders of money, or, to put it differently, it is a reflection of their decisions about the real quantity of money that they desire to hold. We can therefore speak more or less interchangeably about decisions of holders of money to change their real stock of money or to change the ratio of the flow of income to the stock of money.

The situation is more complicated for the monetary arrangements that actually prevailed over the period which our data cover. During part of the period, when the United States was on an effective gold standard, an attempt by holders of money to reduce their cash balances relative to the flow of income raised domestic prices, thereby discouraging exports and encouraging imports, and so tended to increase the outflow of gold or reduce its inflow. In addition, the rise in domestic prices raised, among other things, the cost of producing gold and hence discouraged gold production. Both effects operated to reduce the nominal supply of money. Conversely, an attempt by holders of money to increase their cash balances relative to the flow of income tended to increase the nominal supply of money through the same channels. These effects still occur but can be and typically are offset by Federal Reserve action.

Throughout the period, more complicated reactions operated on the commercial banking system, sometimes in perverse fashion. For example, an attempt by holders of money to reduce cash balances relative to income tended to raise income and prices, thus promoting an expansionary atmosphere in which banks were generally willing to operate on a slenderer margin of liquidity. The result was an increase rather than a reduction in the nominal supply of money. Similarly, changes in the demand for money had effects on security prices and interest rates that affected the amount of money supplied by the banking system. And there were further effects on the actions of the Federal Reserve System for the period since 1914.

There were also indirect effects running in the opposite direction, from changes in the conditions of supply of money to the nominal quantity of money demanded. If, for whatever reason, money-creating institutions expanded the nominal quantity of money, this could have effects, at least in the first instance, on rates of interest and so on the quantity of money demanded, and perhaps also on money income and real income.

Despite these qualifications, all of which would have to be taken into account in a complete analysis, it seems useful to regard the nominal quantity of money as determined primarily by conditions of supply, and the real quantity of money and the income velocity of money as determined primarily by conditions of demand. This implies that we should examine the demand side for an initial interpretation of the observed behavior of velocity.

Along these lines, the changes in the real stock of money and in the income velocity of circulation reflect either (*a*) shifts along a relatively fixed demand schedule for money produced by changes in the variables entering into that schedule; (*b*) changes in the demand schedule itself; or (*c*) temporary departures from the schedule, that is, frictions that make the actual stock of money depart from the desired stock of money. The rest of this paper is an attempt to see to what extent we can reconcile the secular and cyclical behavior of velocity in terms of *a* alone without bringing in the more complicated phenomena that would be involved in *b* and *c*.

One way to do so would be to regard the cyclical changes in velocity as reflecting the influence of variables other than income. In order for this explanation to be satisfactory, these other variables would have to exert an influence opposite to that of income and also be sufficiently potent to dominate the movement of velocity. Our secular results render this implausible, for we there found that income appeared to be the dominant variable affecting the demand for real cash balances. Moreover, the other variables that come first to mind are interest rates, and these display cyclical patterns that seem most unlikely to account for the sizable, highly consistent, and roughly synchronous cyclical pattern in velocity. Long-term corporate interest rates fairly regularly reached their trough in mid-expansion and their peak in mid-contraction prior to World War I. Since then, the pattern is less regular and is characterized by shorter lags. Rates on short-term commercial paper also tend to lag at peaks and troughs, though by a briefer interval, and the lag has similarly shortened since 1921. Call-money rates come closer to being synchronous with the cycle, and this is true also of yields on long- and short-term government obligations for the six cycles for which they are available. Of the rates we have examined these are the only ones that have anything like the right timing pattern to account for the synchronous pattern in velocity. However, neither call-money rates nor government bond yields have been highly consistent in behavior from cycle to cycle. Even if they had been, it seems dubious that the effects of changes in these particular rates, or other unrecorded rates like them, would be sufficiently more important cyclically than secularly to offset the effects of countermovements both in other rates and in income.

Furthermore, earlier studies that have attempted to explain velocity movements in these terms have had only limited success.⁴

A very different way to reconcile the cyclical and secular behavior of velocity is to regard the statistical magnitude called "real income" as corresponding to a different theoretical construct in the cyclical than in the secular analysis. This possibility was suggested by my work on consumption. In that field, too, it will be recalled, there is an apparent conflict between empirical findings for short periods and long periods: cross-section data for individual years suggest that the average propensity to consume is lower at high-income levels than at low-income levels; yet aggregate time-series data covering a long period reveal no secular decline in the average propensity to consume with a rise in income. It turned out that this conflict could be reconciled by distinguishing between "measured" income, the figure recorded by statisticians, and "permanent" income, a longer-term concept to which individuals are regarded as adjusting their consumption.⁵

According to the permanent income hypothesis, when a consumer unit experiences a transitory increment of income, that is, when its measured income exceeds its permanent income, this transitory component is added to its assets (perhaps in the form of durable consumer goods) or used to reduce its liabilities rather than spent on consumption. Conversely, when it experiences a transitory decrement of income, it nonetheless adjusts consumption to permanent income, financing any excess over measured income by drawing down assets or increasing liabilities.

This theory of consumption behavior is directly applicable to that part of the stock of money held by consumer units rather than by business enterprises. The problem is how to interpret money holding. Much of the theoretical literature on "motives" for holding money suggests interpreting money holdings as one of the balance-sheet items that act as shock absorbers for transitory components of income; as an asset item that is increased temporarily when the transitory component is positive and that is drawn down, if necessary, to finance consumption when the transitory component is negative.

This interpretation may be valid for very short time periods. However, if it were valid for periods as long as a business cycle, it would produce a cyclical behavior of velocity precisely the opposite of the observed behavior. Measured income presumably exceeds permanent income at cyclical peaks and falls short of permanent income at cyclical troughs. Hence cash balances would be drawn down abnormally at troughs and

⁴ E.g., see Selden, *op. cit.*, pp. 195–202.

⁵ See my *A Theory of the Consumption Function* (a publication of the National Bureau of Economic Research) (Princeton: Princeton University Press, 1957).

built up abnormally at peaks. In consequence, cash balances would fluctuate more widely over the cycle than income, and velocity would conform inversely to the cycle, falling during expansions and rising during contractions, whereas in fact it conforms positively.

An alternative is to interpret money as a durable consumer good held for the services it renders and yielding a flow of services proportional to the stock, which implies that the shock-absorber function is performed by other items in the balance sheet, such as the stock of durable goods, consumer credit outstanding, personal debt, and perhaps securities held. On this interpretation, the quantity of money demanded, like the quantity of consumption services in general, is adapted not to measured income but to permanent income. This interpretation is consistent with our secular results. The income figure we used in obtaining these is an average value over a cycle, which may be regarded as a closer approximation to permanent income than an annual value. In any case, the long time period covered assures that the movements in money are dominated by the movements in the permanent component of income.⁶ For the cyclical analysis, permanent income need not itself be stable over a cycle. It may well rise during expansions and fall during contractions. Presumably, however, it will rise less than measured income during expansions and fall less during contractions. Hence, if money holdings were adapted to permanent income, they might rise and fall more than in proportion to permanent income, as is required by our secular results, yet less than in proportion to measured income, as is required by our cyclical results.

To put the matter differently, suppose that the demand for real cash balances were determined entirely by real permanent income according to the relation estimated in the secular analysis and that actual balances throughout equaled desired balances. Velocity would then fall during expansions and rise (or fall at a smaller rate) during contractions, *provided* that it was computed by dividing *permanent income* by the stock of money. But the numbers we have been calling "velocity" were not computed in this way; they were computed by dividing measured income by the stock of money. Such a *measured* velocity would tend to be lower than what we may call *permanent* velocity at troughs, because measured income is then lower than permanent income and would tend to be higher at peaks, because measured income is then higher than permanent income. Measured velocity might therefore conform positively to the cycle, even though permanent velocity conformed inversely.

These comments apply explicitly only to consumer cash balances. However, they can readily be extended to business cash balances. Businesses hold cash as a productive resource. The question is whether cash

⁶ *Ibid.*, pp. 125–29.

is a resource like inventories, in which case it might be expected to fluctuate more over the cycle than current production, or like fixed capital, in which case it might be expected to fluctuate less and to be adapted to the longer-term level of production at which a firm plans to operate. This latter possibility involves a concept analogous to that of permanent income. If the observed positive cyclical conformity of velocity reflects wider movements in income than in both business holdings and consumer holdings, as seems likely in view of the changing importance of these two components and the consistent behavior of velocity, the answer must be that cash balances are analogous to fixed capital rather than to inventories and that some other assets or liabilities serve as shock-absorbers for business as for consumers.

The distinction between permanent and measured income can rationalize the observed cyclical behavior of income velocity in terms of a movement along a stable demand curve. It cannot by itself easily rationalize the behavior of real cash balances. Our secular analysis implies that real cash balances should conform positively to the cycle with an amplitude nearly twice that of permanent real income. Observed real cash balances do conform positively, but their amplitude, at any rate for cycles containing mild contractions, is so small that it seems implausible to regard it as larger than that in permanent real income. Put differently, it would take only very moderate changes in the index of prices, well within the margin of error in such indexes, to convert the positive conformity into inverted conformity.

The resolution is straightforward. We have not yet carried our logic far enough. If applied to both money income and real income, the distinction between measured and permanent income implies a corresponding distinction for prices. To put the matter in terms of economics rather than arithmetic, our analysis suggests that holders of cash balances determine the amount to hold in light of their longer-term income position rather than their momentary receipts—this is the justification for distinguishing measured from permanent income. By the same token, they may be expected to determine the amount of cash balances to hold in light of longer-term price movements—permanent prices, as it were—rather than current or measured prices. Suppose, for example, prices were to double permanently or, alternatively, to double for day X only and then return to their initial level and that this behavior was correctly anticipated by holders of money. Holders of money would hardly want to hold the same nominal cash balances on day X in these two cases, even though prices were the same on that day. More generally, whatever the motives for holding cash balances, they are held and are expected to be held for a sizable and indefinite period of time. Holders of money presumably judge the “real” amount of cash balances in terms of the quantity of goods and services to which the

balances are equivalent, not at any given moment of time, but over a sizable and indefinite period; that is, they evaluate them in terms of "expected" or "permanent" prices, not in terms of the current price level. This consideration does not, of course, rule out some adjustment to temporary movements in prices. Such movements offer opportunities of profit from shifting wealth from cash to other forms of assets and conversely, and they may affect people's expectations about future price levels. Like "permanent income," the "permanent" price level need not be—and presumably is not—a constant over time; it departs from the current price level in having a smoother and less fluctuating pattern in time but need not go to the extreme of displaying no fluctuations.

On this view, the current price level would presumably fall short of the permanent price level at troughs and exceed it at peaks of cycles; hence measured real cash balances would tend to be larger than permanent real cash balances at troughs and smaller at peaks. It follows that measured real cash balances would show a smaller cyclical movement than permanent real cash balances and, indeed, might conform inversely to the cycle, even though permanent real cash balances conformed positively.

III. A Symbolic Restatement

The distinction between permanent and measured magnitudes can thus reconcile the qualitative behavior during reference cycles of both measured velocity—its tendency to conform positively—and measured real cash balances—its tendency to show an exceedingly mild cyclical movement—with their behavior over secular periods. The crucial question remains whether it not only can reconcile the qualitative behavior but does in fact rationalize the quantitative behavior of these magnitudes. After all, an interpretation in terms of interest rates can also rationalize the qualitative results; we reject it because it appears likely to be contradicted on a more detailed quantitative level.

It will facilitate such a quantitative test to restate symbolically and more precisely the explanation just presented. Let

- Y be measured aggregate income in nominal terms;
- P be measured price level;
- M be aggregate stock of money in nominal terms,
measured and permanent being taken throughout as identical;
- N be population, measured and permanent being
taken as identical;
- Y_p, P_p be permanent nominal aggregate income and permanent price level, respectively;

$$\begin{aligned}
y &= \frac{Y}{P} && \text{be measured aggregate income in real terms;} \\
y_p &= \frac{Y_p}{P_p} && \text{be permanent aggregate income in real terms;} \\
m &= \frac{M}{P} && \text{be measured aggregate stock of money in real} \\
&&& \text{terms;} \\
m_p &= \frac{M_p}{P_p} && \text{be permanent aggregate stock of money in real} \\
&&& \text{terms;} \\
V &= \frac{Y}{M} = \frac{y}{m} && \text{be measured velocity;} \\
V_p &= \frac{Y_p}{M_p} = \frac{y_p}{m_p} && \text{be permanent velocity.}
\end{aligned}$$

In these symbols, the demand equation fitted to the secular data can be written thus:

$$\frac{M}{NP_p} = \gamma \left(\frac{Y_p}{NP_p} \right)^\delta, \quad (1)$$

which expresses permanent real balances per capita as a function of permanent real income per capita, or in the equivalent form,

$$m_p = \gamma N \left(\frac{y_p}{N} \right)^\delta = \gamma N^{1-\delta} y_p^\delta, \quad (2)$$

which expresses aggregate permanent real balances as a function of aggregate permanent real income and population, where γ and δ are parameters and δ was estimated to be approximately 1.8.⁷

By definition,

$$m = \frac{M}{P} = \frac{MP_p}{P_p P} = \frac{P_p}{P} m_p, \quad (3)$$

so that still a third form of the demand equation is

$$m = \frac{P_p}{P} \gamma N^{1-\delta} y_p^\delta, \quad (4)$$

which expresses aggregate measured real balances as a function of aggregate permanent real income, population, and permanent and measured prices.

This relation can also be expressed in terms of velocity. By definition,

⁷ The basic analysis holds, of course, whatever the precise form of the demand equation for money. I use this particular form for simplicity and because it gave a satisfactory fit to the available evidence. The whole analysis could, however, be restated in terms of a generalized demand function whose form was unspecified.

$V_p = y_p/m_p$. Divide y_p successively by the two sides of equation (2). This gives

$$V_p = \frac{y_p}{m_p} = \frac{1}{\gamma} N^{\delta-1} y_p^{1-\delta} = \frac{1}{\gamma} \left(\frac{y_p}{N} \right)^{1-\delta}. \quad (5)$$

By definition,

$$V = \frac{Y}{M} = \frac{Y Y_p}{Y_p M} = \frac{Y}{Y_p} V_p, \quad (6)$$

so that

$$V = \frac{Y}{Y_p} \frac{1}{\gamma} \left(\frac{y_p}{N} \right)^{1-\delta}. \quad (7)$$

In interpreting equations (1), (2), (4), (5), and (7), it should be borne in mind that they will not, of course, be satisfied precisely by observed data. In consequence, at a later stage, I shall want to distinguish between observed values of, for example, measured velocity and the value estimated from, say, equation (7).

IV. Tests of the Explanation

It has so far been sufficient to suppose only that the permanent magnitudes introduced—permanent income and permanent prices—fluctuate less over the cycle than the corresponding measured magnitudes. We can clearly go farther and ask how much less the permanent magnitudes must fluctuate in order to account for the quantitative, as well as the qualitative, average behavior of velocity and real cash balances. The answer may provide some internal evidence on the plausibility of the suggested explanation and will also provide a starting point for bringing external evidence to bear.

Consider the data for the mild depression cycles shown in Table 1 and neglect the mild cyclical movements in population, so that aggregate and per capita values can be regarded as interchangeable. If measured and permanent magnitudes were treated as identical, the income elasticity of 1.8 computed from the secular data would convert the 0.57 cyclical movement in real income into a movement of 1.03 in *real* cash balances demanded. The movement of 0.14 in the implicit price index would, in turn, convert this into a movement of 1.17 in *money* cash balances demanded. The actual movement in cash balances is 0.27, or 23 per cent as large. Hence, to reconcile the secular and cyclical results, the cyclical movements in permanent income and permanent prices would each have to be 23 per cent of those in measured income and

measured prices—a result that seems not implausible. For deep depression cycles, the corresponding figure turns out to be 37 per cent, which is equally plausible. Moreover, it seems eminently reasonable that this figure should be larger for deep, than for mild, depression cycles, since the deep depression cycles are longer on the average than the mild depression cycles.⁸

Of course, this test of intuitive plausibility is a weak one. To get a stronger test, we must introduce some independent evidence on the relation of permanent to measured magnitudes. One source of such evidence is the work on consumption that suggested the explanation under test. In deriving a consumption function from aggregate time-series data, I concluded that an *estimate* of permanent income—which I called “expected” income to distinguish it from the theoretical concept—was given by

$$y_p(T) = \beta \int_{-\infty}^T e^{(\beta-\alpha)(t-T)} y(t) dt. \quad (8)$$

⁸ Let \dot{M} and \dot{P} be the cyclical movements as measured in the final column of Table 1 in the nominal stock of money and in measured prices; let \dot{m}_p and \dot{P}_p be the cyclical movements in permanent real balances and permanent prices. Then, to a first approximation,

$$\dot{M} = \dot{m}_p + \dot{P}_p \quad (i)$$

since the stock of money is the product of permanent real cash balances and the permanent price level.

Using the demand equation (2), we get

$$\dot{m} = 1.8\dot{y}_p \quad (ii)$$

where \dot{y}_p is the cyclical movement in permanent real income (recall that we are neglecting any cyclical movement in population, so \dot{y}_p also equals the movement in permanent real per capita income).

Let

$$\dot{y}_p = k\dot{y}, \quad (iii)$$

$$\dot{P}_p = k'\dot{P}, \quad (iv)$$

where \dot{y} is the cyclical movement in measured real income and k and k' are unspecified constants to be determined. Substituting equations (ii), (iii), and (iv) in equation (i) gives

$$\dot{M} = 1.8k\dot{y} + k'\dot{P}. \quad (v)$$

At first glance, it seems possible to derive both k and k' from one set of data by deriving a similar equation starting with an identity like (i) expressing measured velocity in terms of permanent velocity. However, the resulting equation is identical with eq. (v), thanks to the definitional relations connecting velocity, money, and income.

The calculations in the text implicitly assume that $k = k'$ in eq. (v). Separate estimates for k and k' require two sets of data. One possibility is to assume that k and k' differ but that each is the same for mild and for deep depression cycles, an assumption that seems less plausible than the one made in the text that $k = k'$. This calculation yields an estimate of 0.11 for k and 1.15 for k' . The value for k' contradicts the concepts of permanent and measured prices that underlie the analysis.

In words, an estimate of expected income at time T is given by a weighted average of past incomes, adjusted for secular growth at the rate of a per cent per year, the weights declining exponentially and being equal to $e^{\beta(t-T)}$, where t is the time of the observation being weighted. The numerical value of β was estimated to be 0.4; of a , 0.02.⁹ It is by no means necessary that the concept of permanent income that is relevant in determining total consumption expenditures should also be the one that is relevant in determining cash balances.¹⁰ But it would not be at all surprising if it were. On the assumption that it is, we can get independent estimates of the percentages cited in the previous paragraph by computing estimates of permanent real income and permanent prices from the corresponding observed annual series, using the weighting pattern just described.

The results of these computations are summarized in columns 1, 2, and 3 of Table 2.¹¹ The agreement between the estimates in column 3 so obtained and the estimates constructed above from internal evidence alone is very good—the two differ by only 15–30 per cent, even though they are based on independent bodies of data and even though the weights used in estimating the permanent magnitudes directly were derived for another purpose and rest on still other data. Moreover, the discrepancy is consistent; the difference between deep and mild depression cycles is in the same direction and of roughly the same magnitude for both columns.

These results are sufficiently encouraging to justify going beyond this indirect test and seeing how far our interpretation is consistent not only with the size of the cyclical movement in cash balances and measured

⁹ Friedman, *A Theory of the Consumption Function*, pp. 146–47.

¹⁰ See *ibid.*, pp. 150–51.

¹¹ These results at first seemed to me relevant also to the choice between the two alternative assumptions used above—the one in the text that $k = k'$ and the one noted in footnote 8, that $k \neq k'$ but that k is the same for mild and deep depression cycles and so is k' . On this issue, the result is unambiguous. The entries in col. 3 clearly speak for the first assumption.

However, James Ford has pointed out to me that this result is largely a consequence of an assumption made in estimating permanent income and prices, namely, the use of the same value of β for both. There is no independent empirical evidence for this assumption, and hence results based on it can give no independent evidence for the essentially equivalent assumption that $k = k'$.

For the special case in which the measured magnitude is given by a sine curve, the relative amplitude of a permanent and a measured magnitude when the permanent is estimated by a weighted average of the measured is determined entirely by the value of β and the duration of the cycle. For $\beta = 0.4$ and a cycle 43 months in length, which is the average length of the mild depression cycles, the relative amplitude for the sine curve is 0.22. For $\beta = 0.4$ and a cycle 47.5 months in length, the average length of the deep depression cycles, the relative amplitude for the sine curve is 0.25. These results are fairly similar to the computed values in Table 2. They differ enough, however, to suggest that the departure from a sine curve affects the results appreciably.

I am indebted to James Ford for these calculations.

TABLE 2
TWO ESTIMATES OF CYCLICAL MOVEMENTS OF PERMANENT REAL INCOME AND PRICES
AS PERCENTAGES OF THOSE OF MEASURED REAL INCOME AND PRICES,
REFERENCE CYCLES 1870-1954, EXCLUDING WAR CYCLES*

	EXCESS OF CHANGE PER MONTH IN REFERENCE- CYCLE RELATIVES DURING REFERENCE EXPANSION OVER THAT DURING REFERENCE CONTRACTION		PERMANENT AS PERCENTAGE OF MEASURED	
	Permanent Magnitude (1)	Measured Magnitude (2)	Permanent Estimated Separately (3)	Estimated from Money Equations (4)
Twelve mild depression cycles:				
Real income	0.11	0.57	19	23
Prices	.02	.14	16	23
Six deep depression cycles:				
Real income	.29	.99	29	37
Prices	0.18	0.60	30	37

* The sources for the columns are as follows (cycles grouped as in Table 1):
1. Permanent real income and permanent prices were estimated as described in the text, using Kuznets' data (see note to Table 1). These data begin in 1869. To obtain an estimate of the permanent magnitude in 1869, measured figures covering the years 1858-69 are required, the weights assigned declining exponentially. Measured figures were therefore extrapolated: for real income by assuming a constant rate of growth of 3.5 per cent per year; for implicit prices by assuming that in each of the years 1858-68 they bore the same relation to the wholesale price index as in 1869.
2. Table 1, col. 3.
3. Column 1 divided by col. 2, the figures in each case being carried to an additional place.
4. Values from Table 1, col. 3, were substituted in the expression $M/(1.82\dot{y} + \dot{P})$, where M is money stock, y is real income, P is implicit price deflator, and the dot on top means "excess of change per month in reference-cycle relatives during reference expansion over that during reference contraction."

velocity but also with their entire cyclical patterns and not only on the average but also cycle by cycle.

In order to perform this test on a fully consistent basis, we first re-computed the secular demand equation, using as the independent variable the cycle averages of estimated permanent income rather than measured income. This substitution slightly raised the correlation coefficient, thus giving a minor bit of additional evidence in favor of the permanent income interpretation. It also raised slightly the estimated elasticity of demand, but not by enough to change the numerical value to the number of significant figures given above.

The resulting calculated equation for nominal cash balances is

$$M^* = (0.00323) \left(\frac{y_p}{N} \right)^{1.810} NP_p \tag{9}$$

and, for measured velocity,

$$V^* = \frac{1}{0.00323} \left(\frac{y_p}{N} \right)^{-0.810} \frac{Y}{Y_p} \tag{10}$$

where the asterisks are used to indicate values computed from the equation rather than directly observed. These equations, it will be recalled, were estimated from average values over whole reference cycles.¹²

From these equations, one can estimate for each year separately, from the corresponding annual data, desired cash balances and the value of measured velocity that would be observed if actual cash balances equaled desired balances as so estimated. I shall call these "computed cash balances" and "computed measured velocity."¹³

The estimates of computed measured velocity are plotted in Chart I, along with observed measured velocity. In judging this figure, it should be borne in mind that the computed velocities were not obtained by trying to fit these observed velocities directly. They were obtained from a correlation for forty-one overlapping cycle bases—averages of groups of years varying in number from two to seven—plus a formula for estimating permanent income derived from an analysis of the relation of consumption expenditures to income plus a theoretical linkage between these two, summarized in equations (9) and (10). The high correlation between the cycle bases insures a close connection between the longer-term movements in computed and measured velocity; in this respect, Chart I is simply a repetition in a different form of the secular finding. What is added by this chart is the relation between year-to-year movements. The secular results in no way insure that these will correspond; still, if anything, the computed velocity series mirrors the year-to-year cycles in observed velocity even more faithfully than it does the longer-term changes.

¹² The numerical values given were computed from combined data for trough-to-trough and peak-to-peak averages. However, separate regressions for each set of averages are almost identical.

¹³ To make these calculations, estimates of Y , Y_p , y_p , P_p , and N are needed. Measured money income, Y , was taken to be Kuznets' annual net national product in current prices adjusted for wartime periods; Y_p was computed by applying eq. (8) to this same series, except for a minor adjustment in level; y_p , by applying eq. (8) to Kuznets' net national product in constant prices similarly adjusted, and again with a minor adjustment in level; P , by applying eq. (8) to the price index implicit in computing net national product in constant prices; and N was taken as the mid-year population of the United States as estimated by the Census.

Equation (8) with $\beta = 0.40$ and $\alpha = 0.02$ implies that expected income is 1.05 times the weighted average of actual income, where the weights are the declining exponential weights inside the integral of eq. (8), adjusted to sum to unity. When permanent net national product per capita in constant prices was computed in this way, it turned out that the geometric mean of the ratios of the cycle bases of real measured net national product per capita to the cycle bases of permanent net national product in constant prices so computed was 1.057. This factor of 1.057 was used to adjust the level of the latter series rather than the 1.05 strictly called for by eq. (8) and was used also for permanent net national product in current prices. The logical implication of employing the same multiple for net national product in constant and current prices is that a was treated as zero for prices alone. None of these adjustments is of any moment for the present analysis, since they affect only the level of the series and hence all cancel out when cycle relatives are computed.

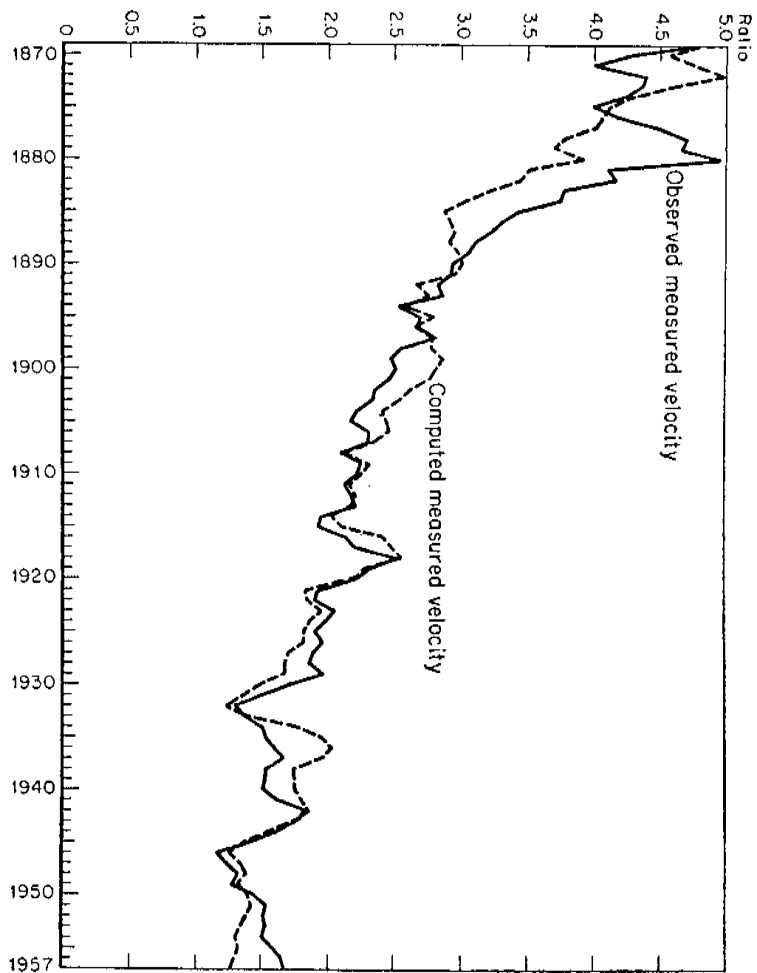


CHART I.—Observed and computed measured velocity, annually, 1869–1957.

In order to isolate the cyclical aspect of the analysis, we have computed reference-cycle patterns of computed measured velocity and computed cash balances, thereby eliminating entirely the part of Chart I that repeats the secular finding. Chart II gives the reference-cycle patterns of computed and observed measured velocity cycle by cycle, and Chart III gives average patterns for the mild and deep depression cycles, for both cash balances and measured velocity. It is clear from these that my interpretation accounts for the bulk of the fluctuations in observed measured velocity. The average pattern of computed measured velocity duplicates almost perfectly that for observed measured velocity for the mild depression cycles and corresponds very closely to that for the deep depression cycles. The cycle-by-cycle patterns demonstrate that this coincidence is not simply in the averages. This closeness might reflect the use of the same values of measured income in both the observed and the computed velocities, in which case it could be regarded as largely spurious. The cash-balance patterns are included in Chart III to test this possibility. They demonstrate that this purely statistical interpretation of the findings is not valid. The cash-balance patterns agree about as closely as the velocity patterns.

These results give strong support to the view that cyclical movements in velocity largely reflect movements along a stable demand curve for money and that the apparent discrepancy between the secular and the cyclical results reflects a divergence between measures of income and of prices constructed by statisticians for short periods and the magnitudes to which holders of money adjust their cash balances.

V. Limitations of the Explanation

Important though this explanation is, it cannot be the whole of the story, since it fails to account for some of the most important of our findings about the behavior of money balances. If the desired real stock of money were determined entirely by permanent real income and if the desired stock were always equal to the actual stock, then the actual real stock (computed in terms of permanent prices) would have a cyclical pattern that duplicated the pattern of permanent real income except for amplitude. Now our evidence suggests that permanent real income conforms positively to the cycle and is either synchronous or lags at the turning points. Hence real cash balances computed at permanent prices would do likewise. Nominal cash balances equal these real cash balances times permanent prices, and our evidence suggests equally that permanent prices conform positively to the cycle either synchronously or with a lag. This train of reasoning therefore implies that, under the supposed conditions, nominal cash balances would conform positively to the cycle and would be either synchronous or lag at

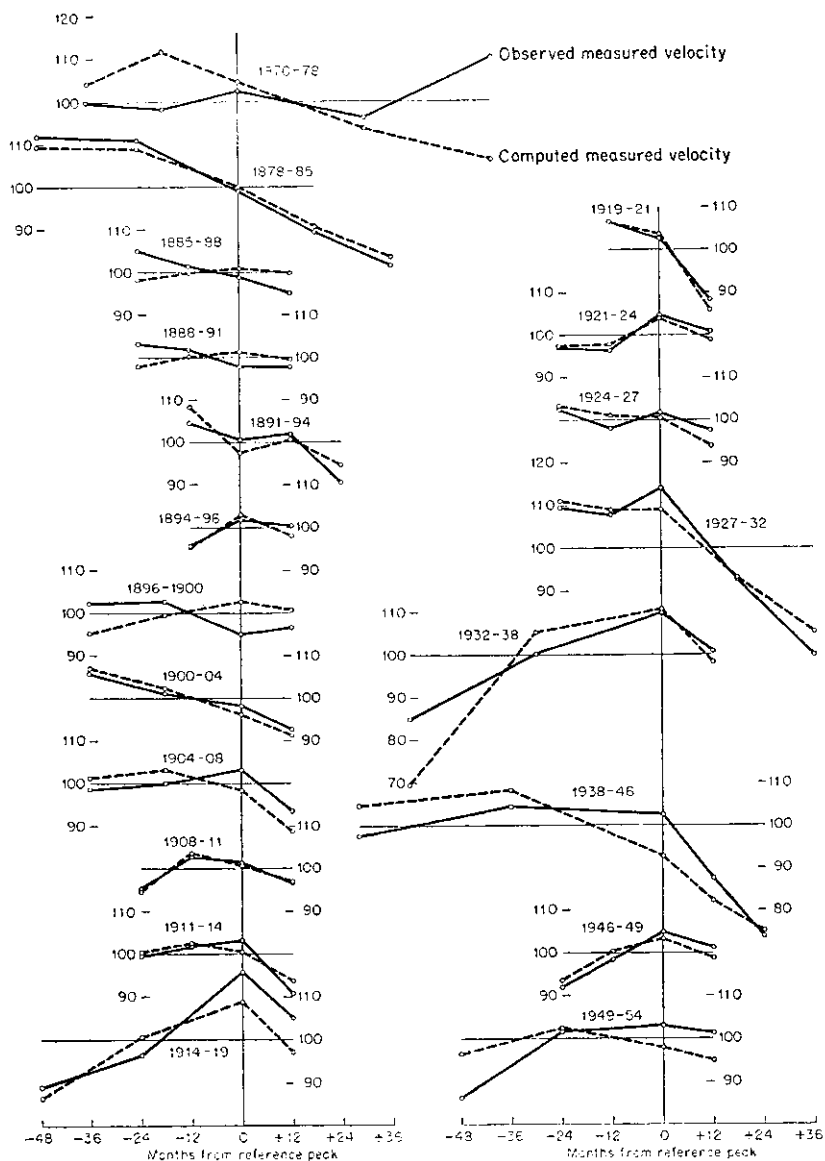


CHART II.—Observed and computed measured velocity, reference-cycle patterns, 1870-1954.

NOTE: These are reference-cycle relatives computed in the course of the cyclical analysis of the data shown in Chart I (see A. F. Burns and W. C. Mitchell, *Measuring Business Cycles* [New York: National Bureau of Economic Research, 1946], pp. 197-202).

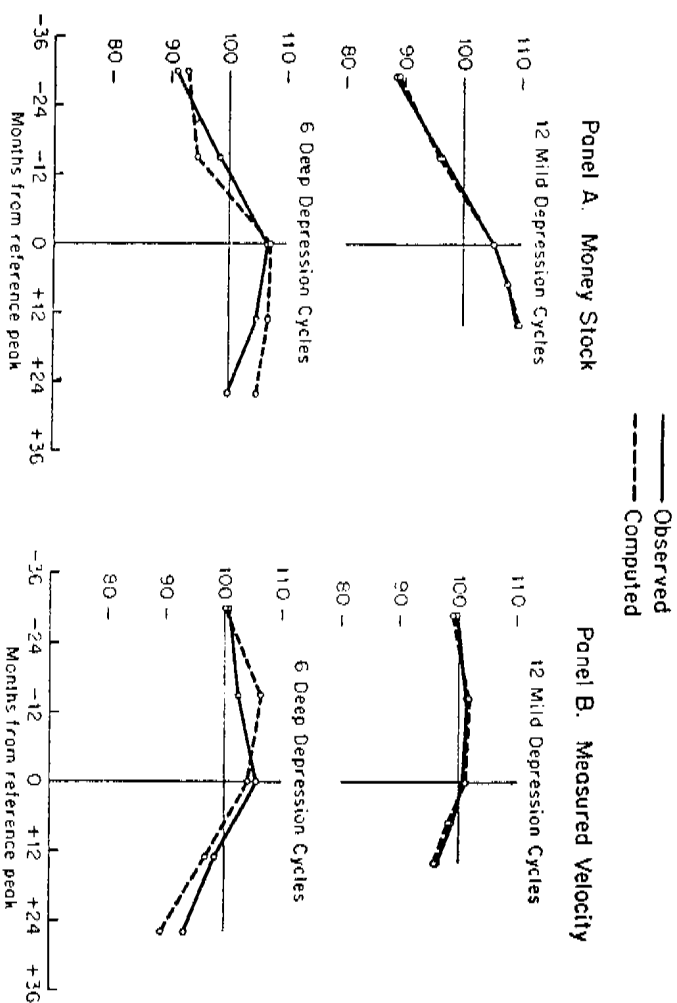


CHART III.—Observed and computed money stock and measured velocity, average reference-cycle patterns, mild and deep depression cycles, 1870–1954.

NOTE: Cycles are grouped as in Table 1.

the turning points. Yet one of the major findings of the broader study of which the results reported in this paper are a part is that the nominal stock of money, adjusted for trend, tends to lead at both peaks and troughs. Hence there is a residual element in the cyclical behavior of velocity that requires explanation.

A satisfactory analysis of this residual element requires the use of monthly rather than annual data. Annual data are unduly crude for studying timing relationships. For example, the cyclical patterns of the observed money stock in Chart III, Panel A, reveal no average lead; yet our more detailed analysis of monthly money data establish such a lead, after adjustment for trend, beyond any reasonable doubt.

It may nevertheless be worth examining the residual element in the annual data as a first step. This residual element is approximated in Chart IV by the ratio of the observed measured velocity to computed measured velocity. This ratio varies very much less over the cycle than measured velocity itself, and hence the movements it measures tend to be concealed by the movements in velocity arising out of the discrepancy between measured and permanent income. Yet our analysis of the stock of money suggests that this residual element may play a critical cyclical role. Indeed, perhaps the major significance of our analysis of velocity is that it enables us to extract this residual element, to eliminate the largely spurious movements of velocity that have hitherto masked the economically significant movements.

For deep depressions, the residual element has a clearly marked cyclical pattern. During expansion, the residual element at first falls, then rises, reaching a trough in mid-expansion. During contractions, the behavior is harder to determine, because one cycle—the earliest, from 1870 to 1878—has a major influence on the pattern for all cycles and the figures for this cycle are highly dubious.¹⁴ If this cycle is omitted, the pattern for contractions is a mild fall from peak to mid-contraction and a sharper fall thereafter.

The residual element varies much less, on the average, for mild depression cycles than for deep depression cycles. Such cyclical movement as it does show is similar to that for deep depression cycles during expansion and just the reverse of that for deep depression cycles during contraction. This residual element is the cyclical component in cash balances that cannot be explained simply by a movement along a univariate demand curve in response to a cyclical movement in permanent income. It is perhaps not surprising that this component should be so

¹⁴ The problem is in the income estimates for the early period. These are characterized by an extraordinarily rapid rate of increase from 1869 to 1879. Other evidence suggests that this is at least partly a statistical artifact, reflecting the extreme paucity of reliable data for estimating income for this period.

much larger for deep than for mild depression cycles. In the mild depression cycles, there is a relatively small cyclical movement in general, which presumably means that there are only relatively small movements in whatever other variables operate to produce a discrepancy between desired cash balances as judged from income alone and actual cash balances.

What are these other variables? The obvious candidates are measures of the return on other assets that could be held instead of money. One alternative to holding money is to hold securities; another, to hold physical goods. The return to the first is measured by the rate of return received on the securities. The return to the holding of physical goods is measured by the rate of change of prices minus storage costs; and either of these terms may be positive or negative—prices may rise or fall and storage of goods may yield a convenience return in excess of costs of handling and maintenance. In either case, these returns must be compared with those on money, which may be positive, as when interest is paid on deposits, or negative, as when service charges are incurred.

In our secular analysis, we have found that the yield on corporate bonds is correlated with the real stock of money and velocity in the expected direction: a rise in the bond yield tends to reduce the real stock of money demanded for a given real income—that is, to raise velocity—and conversely. Bond yields, however, play nothing like so important and regularly consistent a role in accounting for changes in velocity as does real income. The short-term interest rate was even less highly correlated with velocity than the yield on corporate bonds.

Chart IV is designed to provide a rough test whether these secular results carry over to cyclical movements. In addition to the ratio of observed measured velocity to computed measured velocity, which is the residual element we are seeking to explain, Chart IV also shows the average reference-cycle patterns of corporate bond yields as derived from annual data, of commercial paper rates as derived from monthly data (Panel A),¹⁵ and of the yields on short- and long-term United States

¹⁵ The corporate bond yield data through 1900 are railroad bond yields from F. R. Macaulay, *Some Theoretical Problems Suggested by the Movements of Interest Rates, Bond Yields and Stock Prices in the United States since 1856*, a publication of the National Bureau of Economic Research (New York, 1938), pp. A145–A152, col. 5, with 0.114 per cent arithmetic addition to raise them to the level of the following segment. After 1900 the data are “Basic Yields of Corporate Bonds to 50 Years Maturity,” from *Historical Statistics of the United States, 1789–1945* (Bureau of the Census), p. 279; *Continuation to 1952 of Historical Statistics*, p. 36; *Statistical Abstract of the United States*, annually from 1953. Commercial paper rates in New York City, monthly, through January, 1937, are from Macaulay, *op. cit.*, pp. A145–A161; thereafter, monthly averages of weekly figures from *Bank and Quotation Record of the Commercial and Financial Chronicle*. This series was seasonally adjusted through December, 1933. No seasonal adjustment has been necessary since.

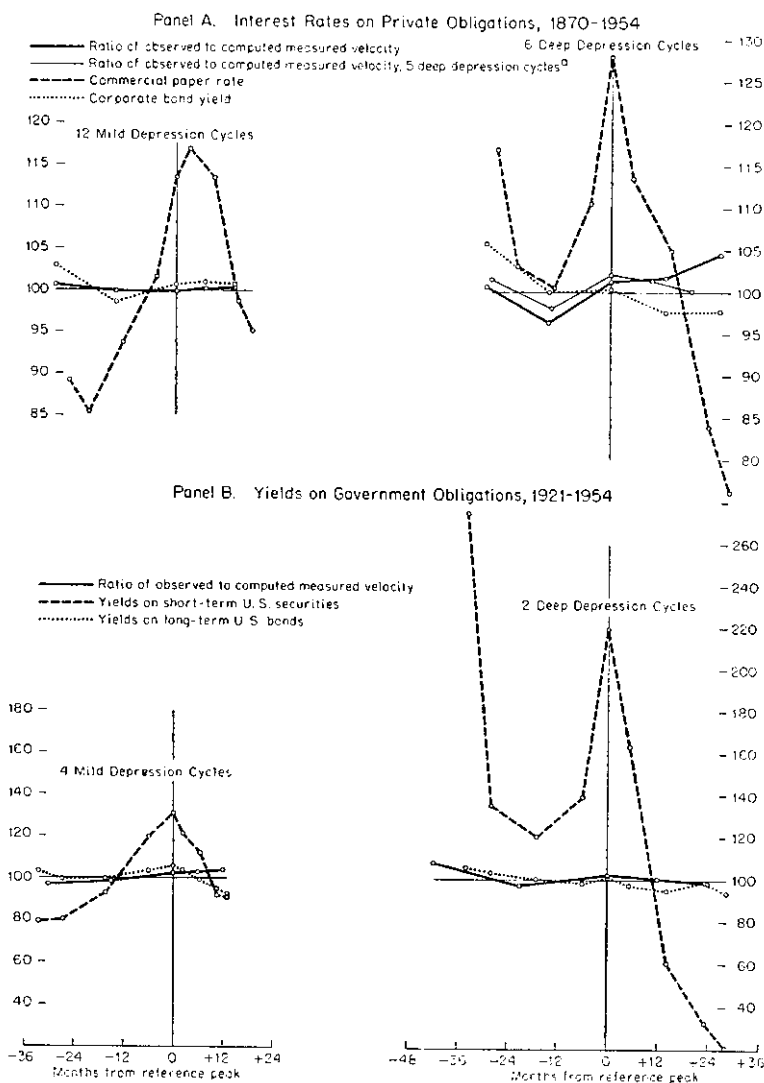


CHART IV.—Ratio of observed to computed measured velocity, compared with other economic variables, average reference-cycle patterns, mild and deep depression cycles, 1870-1954.

NOTE: Vertical scales are in reference-cycle relatives, except scale for prices in Panel C, which is in rate of change of reference-cycle relatives per month. The scale of reference-cycle relatives in Panel B is one-fourth that in Panel A, and the scale in Panel C is two and a half times that in Panel A.

^a Excluding 1870-78.

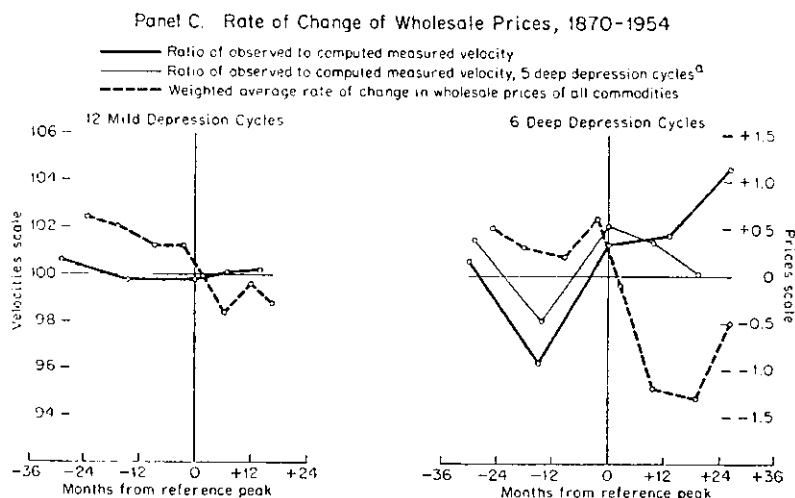


CHART IV.—Continued

securities, as derived from monthly data (Panel *B*).¹⁶ Panel *A* covers the whole period 1870-1954, excluding only war cycles; Panel *B* covers only the six non-war cycles after 1921, since yields on United States securities are not readily available for the earlier cycles.

Short-term rates have, of course, a much larger cyclical amplitude than long-term rates, which in turn have roughly the same amplitude as the residual element in velocity. These differences in amplitude are of no special significance for our purpose except as they reflect the consistency of the cyclical pattern, since the effect of a change in interest rates depends not only on the size of the change but also on the elasticity of the response of cash balances to a change. Volatility of rate can be offset by a small elasticity of response and vice versa. The differences in amplitude do, however, make it more difficult to read the chart and tend somewhat to obscure the similarity or divergence in pattern that is of major interest.

The most striking feature of the charts is the high degree of similarity between the pattern of interest rates and that of the residual element of velocity during the expansion phase of deep depression cycles. Long and short rates and rates on private and public obligations all show much the same pattern for this phase, and the pattern of all four is similar to the pattern in the residual element in velocity: interest rates

¹⁶ Yields on short-term United States securities are from *Banking and Monetary Statistics*, p. 460, and *Federal Reserve Bulletin*, monthly issues, May, 1945, to May, 1948, and September, 1950, to December, 1954. This series was seasonally adjusted, 1920-30, 1951-54. Yields on long-term United States securities are from the same sources and are unadjusted.

are high at the initial stage of expansion, and so is velocity, which is an appropriate response to a high rate of return on non-cash assets; interest rates then decline to mid-expansion, and so does velocity; interest rates then rise to the peak of the cycle, and so does velocity.

There is no such unanimity of movement for the remaining phase of the deep depression cycles or for the mild depression cycles. For these phases, there is, at best, a family similarity between the movements in rates and those in the residual element in velocity. During the contraction phase of deep depression cycles, short and long rates diverge, short rates declining throughout, long rates leveling off or recovering in mid-contraction. The residual element behaves rather more like short rates, if we abstract from the unusual behavior during the 1870–78 cycle, but the similarity is not close in detail. For mild depression cycles, the cyclical movements in short and long rates are fairly similar, the main differences being a shorter lag in commercial paper rates at peaks and troughs than in the corporate bond yield. For the period as a whole (Panel A), the cyclical movement in the residual element, though fairly clear, is so small that no very precise comparison is justified; for the period since 1921 (Panel B), it is almost non-existent, the average reference-cycle pattern being dominated by an intracycle trend.

A number of empirical studies have demonstrated that the rate of change of prices has an important effect on the quantity of money demanded during periods of considerable instability of prices—as during hyperinflations or major and long-continued inflations.¹⁷ These studies suggest, further, that the expected rate of change of prices, which is the variable that directly influences the demand for money, can be regarded as derived largely from past experience with the actual movement of prices and that it changes more smoothly than actual prices; it is something like the rate of change in what I earlier designated “permanent” prices. These findings imply that any changes in the expected rate of change of prices during periods of relative price stability will be small, perhaps too small to have any appreciable effect. And this is, indeed, the conclusion reached by Richard Selden in his study of the behavior of velocity.¹⁸

As a further check on this conclusion, we have plotted in Chart IV, Panel C, the rate of change of prices from reference stage to reference stage. This is derived from the nine-stage reference-cycle patterns of the

¹⁷ See Phillip Cagan, “The Monetary Dynamics of Hyperinflation,” in Milton Friedman (ed.), *Studies in the Quantity Theory of Money*, pp. 25–117. The same relation has been documented for other countries and episodes in a number of unpublished studies done in the Workshop on Money and Banking of the University of Chicago.

¹⁸ Selden, *op. cit.*, p. 202.

monthly wholesale price index,¹⁹ by dividing the difference between successive average standings by the average time interval between them. The resulting eight rates of change per month are plotted at the mid-points of the corresponding intervals. Since these are the actual rates of change, they presumably vary more than expected rates of change and, in addition, may lead the latter in time. However, one might expect enough similarity between the actual rates of change and the expected rates of change to permit the detection of any moderately close relation between expected rate of change and the residual element in velocity.

Interestingly enough, the results largely duplicate those for interest rates. For the expansion phase of the deep depression cycles, there is the same striking agreement in pattern between the rate of change of prices and the residual element in velocity as there is between interest rates and the residual element. There is only slightly less similarity in pattern for the expansion phase of mild depression cycles. There is no systematic relation for the contraction phase of either group of cycles.

This analysis, based as it is on annual velocity data and on a comparison solely of average reference-cycle patterns, is too crude to be at all decisive. Yet the results are most suggestive. If the cyclical patterns of interest rates and the rate of change in prices are compared with the pattern of measured velocity itself (Chart III, Panel *B*), there is no clear relation—as we noted at the outset in explaining why an alternative reconciliation of the secular and cyclical behavior of velocity is required. When the comparison is made instead with the residual element of velocity—that part of the movement in measured velocity that is accounted for neither by the effect of changes in permanent income on desired cash balances nor by the discrepancy between measured and permanent income—there is a striking consistency for one phase of one set of cycles, and at least a family resemblance elsewhere, though, of course, not without considerable irregularity. These results are of the kind that might be expected if the returns on alternative ways of holding assets were the chief factor other than permanent income affecting desired cash balances. Of course, they do not demonstrate that this is so. They might, for example, reflect accidental concurrence of movement in just a few cycles. And they do not provide any estimate of the quantitative strength of the connection. But they certainly justify further research in this direction. The main requirements for such research are the use of monthly data on velocity or indicators of velocity and the examination of cycle-by-cycle relations and not simply relations between average patterns.

¹⁹ *Historical Statistics of the United States, 1789–1945* (Warren-Pearson series, 1870–89; B.L.S. series, 1890–1945 [Bureau of the Census]), p. 344; *Continuation to 1952 of Historical Statistics*, p. 47; thereafter, U.S. Department of Labor, Bureau of Labor Statistics, *Wholesale (Primary Market) Price Index*, monthly issues.

VI. Conclusion

The results summarized in this paper have implications for the theory of money, the study of business cycles, and the conduct and possibilities of monetary policy.

In the theory of money, much emphasis has been placed on different “motives” for holding money—the “transactions” motive, the “speculative” motive, and the “assets” or “precautionary” motive being the three commonly distinguished. The transactions motive is often regarded as implying something of a quasi-mechanical relation between cash balances and the flow of payments and is frequently given priority of importance as well as place. Our results cast serious doubt on the acceptability of this emphasis. In the first place, the cyclical results make it clear that changes in cash balances over short periods are adapted to magnitudes less volatile than the volume of transactions. In the second place, the secular decline in income velocity is hard to explain in terms of transactions. It is dubious that there has been any secular increase in the ratio of transactions to income large enough to explain the growth in the ratio of money balances to income that has occurred. Further, improvements in transportation and communication, let alone in financial organization, have almost surely reduced any mechanical requirement for cash balances per unit of transactions—indeed, it was on these grounds that Irving Fisher implied nearly half-a-century ago that velocity was likely to increase secularly and that others have since expressed similar views.²⁰

Our findings equally cast doubt on the importance of the so-called speculative motive. One would expect this motive to be subject to wide cyclical variations and hence, if it dominated the demand for money, to lead to correspondingly wide cyclical variations in desired cash balances, whereas we observe the reverse.

The assets or “precautionary” motive is in a different state. Permanent income can be regarded as a concept closely allied to wealth and indeed as an index of wealth, provided that we count both human and non-human sources of income as components of total wealth. Along these lines, our results can be interpreted in either of two ways. One is that the relevant asset motive is equivalent to a consumption or income motive. As permanent income, which is to say, total wealth, rises, consumer units expand their expenditures on some items disproportionately—we term these items “luxuries.” On this interpretation, the services rendered by money can be included among these luxuries. The other interpretation is more nearly an asset motive proper. It is that the holdings of cash are linked not to total wealth but primarily to non-

²⁰ Irving Fisher, *The Purchasing Power of Money* (rev. ed.; New York, 1913), pp. 79–88.

human wealth and that, as permanent income rises, the total value of non-human wealth rises more rapidly than permanent income, either because such a more rapid rise is a necessary condition for a rise in income or because it corresponds to the preferences of individuals as their total wealth rises. Unfortunately, the available evidence on the secular or cyclical behavior of the ratio of non-human wealth to income is inadequate to provide a test of this explanation.²¹ On either interpretation, however, our results suggest that motivations and variables linked with assets are the most fruitful category to explore—that the most fruitful approach is to regard money as one of a sequence of assets, on a par with bonds, equities, houses, consumer durable goods, and the like.

Our results have a bearing on another aspect of the so-called precautionary motive, namely, the view that the amount of cash balances held is highly sensitive to “the” or “a” rate of interest, at least for some range of rates of interest. If this were so for rates of interest within the range observed during the period our data cover, it would imply that real cash balances and the ratio of income to money would be highly variable, both secularly and cyclically, since small movements in interest rates would be accompanied by large movements in desired cash balances. The highly stable secular behavior of velocity is evidence against this view. So is our inability to find any close connection between changes in velocity from cycle to cycle and any of a number of interest rates. So also is our finding that most of the cyclical movement in income velocity as ordinarily measured can be accounted for by the use of measured rather than permanent income in the numerator. The remaining movement in velocity, though characterized by a consistent cyclical pattern and though, on the basis of our tentative explorations, it may well be accounted for by movements in interest rates, is much too small to reflect any very sensitive adjustment of cash balances to interest rates.

Some of these comments about the implications of our results for the theory of money have their direct and obvious counterparts for the empirical study of business cycles. The most important additional implications are two that have to do with the interpretation of cyclical movements in velocity. The fact that velocity changes have been about as important as changes in the stock of money in accounting, in an arithmetic sense, for the movements in money income, together with the small amplitude of cyclical movements in the stock of money, has fostered the view that changes in the stock of money cannot be the prime mover, or even of major independent importance, in cyclical

²¹ Raymond Goldsmith's estimates in *A Study of Savings* (Princeton, N.J., 1955) suggest that, if anything, the ratio of non-human wealth to income has declined secularly rather than risen.

change. This view may of course be correct, but it needs re-examination in light of our finding that most of the velocity movement is, from one point of view, "spurious," as well as a possible consequence of this finding, discussed more fully below, that measured income may be highly sensitive to changes in the stock of money. The other important implication for the study of cycles is that the cyclical pattern of velocity changes that needs study and explanation is very different from what it has been supposed to be. Measured velocity has a cyclical pattern roughly synchronous with that in general business, tending to rise relative to its trend from reference trough to reference peak and to fall from reference peak to reference trough. But when this pattern is corrected for the deviation of measured income from permanent income, the residual movement is very different, and it is the residual movement that needs explanation.

The most interesting implication of our analysis for monetary policy is highly speculative and involves taking our findings more seriously in detail than I can fully justify. It may nonetheless be worth recording if only in the hope of stimulating further work. Suppose one accepts fully both the reasonably well-supported finding that money holdings are adapted to permanent magnitudes and also the much more questionable and tentative suggestion that the economic actors derive their estimates of permanent magnitudes from prior measured magnitudes by implicitly constructing some kind of weighted average of them. It will then follow that, given a stable demand function for money, measured income will be highly sensitive in short periods to changes in the nominal stock of money—the short-run money multiplier will be large and decidedly higher than the long-run money multiplier.²² To illustrate with some figures based on our tentative results: In the long run, if we take real income as given, a \$1 increase in the stock of money would imply an annual level of money income higher than otherwise by \$1 times the velocity of circulation, or, at current levels of velocity, about \$1.50 higher—the long-run money multiplier equals the velocity of circulation. In the short run, however, an increase of \$1.50 in measured income would be inadequate, since that much of a rise in measured income would raise permanent money income by decidedly less than \$1.50 and hence desired cash balances by less than \$1. If we take a year as our unit and accept the numerical weights we have used in estimating permanent income from measured income, measured income would have to rise by roughly \$4.50 for estimated permanent income to rise by \$1.50, the rise required to raise desired cash balances by \$1 for given real income—the short-run money multiplier is thus triple the long-run multiplier.

²² This point was first suggested to me by Gary S. Becker.

The story does not, of course, end here. There would be carry-over effects into future years, as estimated permanent income continued to be revised in the light of measured income. These would make the initially assumed rise in money income not sustainable without further rises in the stock of money and hence would give rise to a cyclical reaction in measured income. Further, the assumed change in money income would presumably be associated with changes in output and in prices that would affect the relation of desired cash balances to the change in measured money income. These further complications require much more study than I have given them. They do not, however, affect the main point—the sensitivity of measured income to changes in the stock of money that is implied by our results if they are accepted at face value.

It is interesting that the permanent-income hypothesis should have such contrasting implications for the sensitivity of the economy to changes in the stock of money and to changes in investment—the major other factor regarded as a prime mover in cyclical change. The permanent-income hypothesis implies that the economy is much less sensitive to changes in investment than it would be if consumption were adapted to measured rather than permanent income—the short-run investment multiplier is decidedly smaller than the long-run multiplier.²³ On the other hand, we have just seen that the economy is much more sensitive to changes in the stock of money than it would be if money balances were adapted to measured rather than permanent income.

A corollary for policy is that the effects of monetary policy may be expected to operate rather more than would otherwise be supposed through the direct effects of changes in the stock of money on spending, and rather less through indirect effects on rates of interest, thence on investment, and thence on income. Another corollary is to emphasize the potency of relatively small changes in the stock of money—a potency, needless to say, for good or evil. Relatively small changes in the stock of money, properly timed and correct in magnitude, may be adequate to offset other changes making for instability. On the other hand, relatively small changes in the stock of money, random in timing and size, may equally be an important source of instability. If the reaction mechanism I have described is in any substantial measure valid, the system may not have a large tolerance for mistakes in monetary management.

²³ See *A Theory of the Consumption Function*, p. 238.

Interest Rates and the Demand for Money

One major strand of Keynesian analysis traces the implications of a particular empirical assumption about the demand for money—that its elasticity with respect to interest rates is very high, approaching infinity (in Keynes' own terms, liquidity preference is, if not absolute, approximately so). Such a situation would have very far-reaching implications: it would greatly limit the effectiveness of price flexibility in correcting unemployment; it would render changes in the quantity of money produced by open market operations impotent to affect economic conditions; it would make the effect of government deficits on income and employment independent of the way in which the deficits are financed.

By now, there is wide agreement that conditions of near-absolute liquidity preference, if they occur at all, are very rare, so that this strand of Keynesian analysis has receded to the status of a theoretical curiosity.

More recently, a number of economists have attributed major theoretical importance to the opposite empirical assumption about the demand for money—that its elasticity with respect to interest rates is negligible. Such a situation, they assert, would have far-reaching implications for the theoretical possibility of separating monetary and real forces and for monetary policy.

Like Keynes' analysis, these assertions raise two separable issues. One issue is empirical: What is the elasticity of the demand for one or another monetary total with respect to various interest rates? How stable is the relation between the (real) quantity of money demanded and interest rates, for both different monetary totals and different interest rates? How consistent are the elasticities for different periods and countries? How important are interest rates compared with other variables in explaining changes in the quantity of money demanded? The other issue

I am grateful to David Fand, David Lindsey, and George Tolley for helpful comments on an earlier draft of this paper. I was helped also by a general discussion at the Workshop in Money and Banking at the University of Chicago, and by extensive earlier correspondence with Harry G. Johnson on some of the main issues.

is theoretical: Would a highly inelastic demand for money with respect to interest rates have the far-reaching implications alleged?

There already is something of a consensus on the empirical issue, though the natural tendency for writers to differentiate their products tends somewhat to conceal it. (1) I know no empirical student of the demand for money who denies that interest rates affect the real quantity of money demanded—though others have misinterpreted me as so asserting.¹ (2) There is no agreement whether short-term or long-term interest rates are more closely related to the quantity of money demanded, though it is clear that elasticities are lower (in absolute value) for short than for long rates. (3) Almost all estimates, even for long

¹ This misunderstanding stems from my article, *The Demand for Money: Some Theoretical and Empirical Results*, 67 J. Pol. Econ. 327 (1959) (reprinted as Occasional Paper No. 68 of the National Bureau of Economic Research (1959)). The empirical demand function presented in that article did not include interest rates as a variable. Further, in summarizing my conclusions, I stated “our inability to find any close connection between changes in velocity and any of a number of interest rates” is evidence against “the view that the amount of cash balances held is highly sensitive to ‘the’ or ‘a’ rate of interest.” *Supra* at 349.

However, inability to pin down the elasticity is very different from assigning a zero value to it. Neither in that article, nor, to the best of my knowledge, elsewhere, have I ever asserted that interest rates have no effect on the quantity of money demanded or on velocity, only that (a) they appear to be less important as a determinant of quantity demanded than real per capita income and as a determinant of measured velocity than the ratio of measured to permanent income; and (b) that the interest elasticity is not very high. Both of these conclusions have on the whole been supported by subsequent research.

I stressed at a number of points in that article the potential significance of interest rates. In answer to the question what variables other than income and prices affect the quantity of money demanded, I wrote:

The obvious candidates are measures of the return on other assets that could be held instead of money . . . , the rate of return received on . . . securities . . . [and] the rate of change of prices. . . .

In our secular analysis, we have found that the yield on corporate bonds is correlated with the real stock of money and velocity in the expected direction. . . . Bond yields, however, play nothing like so important and regularly consistent a role in accounting for changes in velocity as does real income. The short-term interest rate was even less highly correlated with velocity than the yield on corporate bonds. *Supra* at 345.

I then went on to make a “rough test whether these secular results carry over to cyclical movements,” and subsequently, also, a rough test of the effect of the rate of change of prices. My conclusion was

This analysis, based as it is on annual velocity data and on a comparison solely of average reference-cycle patterns is too crude to be at all decisive. Yet the results are most suggestive. . . . These results are of the kind that might be expected if the returns on alternative ways of holding assets were the chief factor other than permanent income affecting desired cash balances. Of course they do not demonstrate that this is so. . . . But they certainly justify further research in this direction. *Supra* at 347–348.

It is baffling to me how anyone could interpret this statement as asserting that interest rates have no effect on the quantity of money demanded!

In chapter 12 of Friedman and Schwartz, *A Monetary History of the United States, 1867–1960* (1963), we used an estimate of the interest-elasticity of money of about $-.15$. This is lower in absolute value than the elasticity estimated by others, and indeed by ourselves in later work. But that is simply a question of different estimates of an empirical magnitude.

rates, show an inelastic response, i.e., elasticities less than unity in absolute value, though most estimates including some we have obtained in our own subsequent work, are higher (in absolute value) than the estimate Anna J. Schwartz and I used in *A Monetary History* (−.15). (4) With only one exception, every study for the United States I know of finds that variations in real income or wealth are a more important source of variations in the real quantity of money demanded than are variations in interest rates.² (5) There is less agreement about the relative importance of different variables for velocity than for the real quantity of money demanded. Some studies find income or wealth elasticities around unity for some monetary totals, which means that the corresponding velocities are independent of income. In many of these studies, interest rates are treated as the major variable affecting velocity. Other studies find income or wealth elasticities significantly different from unity and hence find income or wealth variables as important as or more important than interest rates in explaining variations in velocity, and some find still other variables dominating velocity.³

Since empirical work on the demand for money is a dramatic growth industry, additional evidence on the empirical issues is accumulating rapidly and we may confidently expect a still further convergence of answers.

The promise of a consensus on the empirical issue renders it all the more important to examine the theoretical issue. Are “fundamental issues in monetary theory” associated with the precise answers reached, as one writer has stated?⁴

I believe that only a finding of near-absolute liquidity preference would raise such fundamental issues, and that any other finding would not. It is important to determine as accurately as possible the size of the elasticities in order to have a better empirical basis for understanding the course of economic events and for guiding policy. The size of the elasticities will have important effects on the quantitative magnitude of changes in certain economic variables that can be expected to be produced by changes in other economic variables. But the precise value of

² The reason for limiting this statement to the U.S. studies is because in countries that have experienced substantial or hyper-inflation, the rate of change of prices is generally a much more important source of variation than real income.

For the United States, the one exception I know of is a study of post-World-War II, U.S. quarterly data for households by Hamburger, *The Demand for Money by Households, Money Substitutes, and Monetary Policy*, 74 J. Pol. Econ. 600 (1966).

³ The comment in the preceding footnote for periods of substantial inflation applies here as well. For such periods, anticipated rates of change of prices are almost always the most important single variable affecting velocity. For a recent summary of the evidence, see Laidler, *The Rate of Interest and the Demand for Money—Some Empirical Evidence*, 74 J. Pol. Econ. 543 (1966).

⁴ Johnson, *A Quantity Theorist's Monetary History of the United States*, 75 Economic J. 388, 396 (1965).

the elasticity will not, in my opinion, have major implications for either fundamental issues or the basic role and functioning of monetary policy. The purpose of this note is to explain and justify that conclusion.

The theoretical issue was forcibly impressed on me by two reviewers of *A Monetary History*.⁵ Both reviewers criticized Mrs. Schwartz and me severely for assigning a low interest elasticity to the demand for money. Both asserted that our doing so had far-reaching implications for the conclusions in that book and for policy recommendations I have made elsewhere. However, this note is not intended primarily to reply to their criticisms, though I hope that as a by-product it will do so. After those reviews called the issue to my attention, I came across repeated statements of a similar kind in other connections. Hence, I believe that the issue is of more general significance.

I. The Theoretical Issue

I can best present the issue by quoting the reviewers already referred to.

H. G. Johnson: "[1] If interest rates do not affect velocity, monetary analysis can be divorced from analysis of the real sector, since the quantity of money will affect money income in the short run and prices in the long run without interference from the real forces. If, on the other hand, interest rates do affect velocity, monetary analysis must incorporate the real sector in a general equilibrium model simultaneously explaining interest rates, velocity, real income and prices. [2] Moreover, this need for a general equilibrium model comprising the real and monetary sectors is what the Keynesian Revolution was about; hence to admit interest rates into the demand function for money is to accept the Keynesian Revolution and Keynes' attack on the quantity theory. [3] And, finally, in the absence of a velocity function independent of interest rates, the case for replacing discretionary monetary management by a fixed rule of monetary increase related to the normal growth of the economy, advocated elsewhere by Professor Friedman, loses its attractiveness, because variations in interest rates generated by the real sector would make such a policy rule automatically destabilizing."⁶

Allan Meltzer: "[4] Had the authors systematically incorporated interest rates or asset yields as a determinant of velocity or of the money supply, they would have been forced to do what they have otherwise avoided doing, develop a more extensive analysis of the real system to supplement their treatment of the monetary sector."⁷

⁵ Johnson, *ibid.*; Meltzer, *Monetary Theory and Monetary History*, 101 *Schweizerische Zeitschrift für Volkswirtschaft und Statistik*, 404 (1965).

⁶ Johnson, *id* at 396, numbers added.

⁷ Meltzer, *op. cit. supra* note 5 at 420, numbers added.

I may add one quotation from a non-reviewer, though I should note that it was partly called forth by his discussion of my work.

Daniel Brill: "The significance for the conduct and evaluation of monetary policy of dropping the assumption that money demand is almost completely interest-inelastic deserves more careful attention than it has been given. . . .

"[5] . . . Interest-elasticity of the public's demand for money breaks the tight linkage between the stock of money and money income. It permits fluctuations in propensities to spend, given the money stock, to influence equilibrium interest rates and income; it also [6] allows fiscal policies to alter the level of aggregate expenditures for goods and services, quite apart from their influence on the stock of money.

"[7] The degree of financial restraint or stimulus imposed on the economic system, accordingly, is no longer reflected in any simple way by variations in the money stock. . . ."⁸

II. The Divorce between Monetary and Real Factors

The comments I have numbered [1], [4], and [5] all assert (or imply) that exclusion of interest rates from the demand function for money (zero elasticity) permits (or requires?) a divorce between the monetary sector and the real sector of the economy, whereas inclusion of interest rates (non-zero elasticity) renders such a divorce impossible. As a matter of pure theory these statements seem to me either seriously incomplete or flatly wrong depending on the precise interpretation of the statement "monetary analysis can be divorced from analysis of the real sector."⁹

Two different interpretations seem worth attention:

(a) That knowledge of the nominal quantity of money alone is enough in principle to permit prediction of (that is, to determine) the level of nominal income¹⁰ and perhaps also of prices.¹¹ On this interpretation, the assertion is incomplete: exclusion of interest rates is a

⁸ Brill, *Criteria for the Conduct of Monetary Policy: The Implications of Recent Research*, paper delivered at Conference of University Professors, sponsored by American Bankers Association and Purdue University, on September 1, 1965; numbers added.

⁹ They are also wrong if, as some Keynesians have done, interest rates are themselves regarded as a purely monetary phenomenon, determined by liquidity performance, and having no influence on the real sector through either investment or consumption. However, the writers quoted clearly assume implicitly that interest rates depend on and influence real magnitudes and since I agree fully with them that such an assumption is more useful, I have neglected what might be called the strict Keynesian case.

¹⁰ I am using "nominal income" as a synonym for what Johnson refers to as "money income" in order to avoid using "money" in two different senses.

¹¹ To keep the analysis simple, I am interpreting monetary "analysis" as referring to analysis of the quantity of a single monetary total called money. A more general treatment would recognize a number of monetary magnitudes, such as currency, demand deposits, time deposits, etc. This would complicate the exposition without, so far as I have been able to determine, affecting any essential point.

necessary but not a sufficient condition for a divorce in this sense. However, if "level of nominal income" is replaced by "changes in nominal income," exclusion of interest rates is not even a necessary condition let alone a sufficient one.

(b) That the nominal quantity of money and changes in it have no effects on real magnitudes, including interest rates, although real magnitudes may affect the level of incomes and prices associated with a given nominal quantity of money. In this, which I believe is much the more important sense, the exclusion of interest rates is neither a necessary nor a sufficient condition. It is simply irrelevant.

As a matter of experience, the conditions necessary for a divorce of the first kind are often approximated for changes in nominal income and, less frequently, in prices. I believe that they are seldom if ever approximated for a divorce of the second kind. Indeed, the central message of our *Monetary History* is precisely that monetary changes have an extraordinarily important impact on real phenomena.

Let me turn to a more detailed examination of these two interpretations.

(a) *Does the quantity of money alone determine nominal income and prices?*—If interest rates enter the demand function for money,¹² it is clearly impossible to predict the level of nominal income or of prices solely from the nominal quantity of money. Knowledge of the interest rate, which is to say, indirectly of the real forces affecting the interest rate, is necessary to get a numerical value of velocity. However, if interest rates are stable, knowledge of interest rates is not necessary to predict changes in nominal income or in prices, so exclusion of interest rates is not even a *necessary* condition for a divorce for such magnitudes.

For sufficient conditions, it is best to consider nominal income and prices separately.

For nominal income, a divorce in the sense under discussion between monetary and real factors requires not only that interest rates not enter the demand function but also that (i) no other real factors enter the function giving desired velocity, and (ii) either desired velocity is always equal to actual velocity or the adjustment between the two does not depend on real factors. If these conditions are satisfied, velocity is either a numerical constant or a function solely of the past history of the quantity of money. For changes in nominal income, it is sufficient that any real factors that enter the demand or adjustment functions be stable.

To be more concrete, consider the demand functions for money that have generally been used in recent empirical work. These make the real quantity of money demanded a function of population and real per

¹² I shall throughout interpret this expression as meaning "enter with an elasticity different from zero."

capita income, as well as of interest rates, rate of change of prices, and other variables. Suppose all variables except population and real income per capita are excluded, that the elasticity with respect to population is taken as unity, and that desired real money balances are assumed always equal to actual real money balances. Even then it is necessary for the elasticity of demand for real money balances per capita with respect to real per capita income to be unity in order to divorce monetary analysis from the real sector in the sense under consideration.¹³ This would make velocity independent of real factors.¹⁴

For prices, I have been able to formulate no economically meaningful conditions under which their level would depend only on monetary magnitudes. Even if nominal income is divorced from the real sector in the above sense, the level of prices depends on total real output. However, *changes* in prices could be regarded as purely monetary under

¹³ Let

$$\frac{M^D}{NP} = f\left(\frac{Y}{NP}\right) \quad (1)$$

be a demand function expressing the real per capita balances desired as a function solely of real per capita income, where M^D is desired nominal balances, N is population, P price level, and Y nominal income, and assume that

$$M^D = M^S, \quad (2)$$

where M^S is nominal quantity of money supplied. If (1) is unit elastic, it becomes

$$\frac{M^D}{NP} = k \cdot \frac{Y}{NP}, \quad (1')$$

and we can solve (1') and (2) to get

$$Y = \frac{1}{k} M^S, \quad (3)$$

so knowledge of M^S alone is enough to predict Y . If (1) is not unit elastic, then the counterpart of (3) will take the form

$$Y = g\left(\frac{Y}{NP}\right) \cdot M^S \quad (4)$$

so it is necessary to know real per capita income as well as M^S to predict Y .

Alternatively suppose that (1') holds but that (2) is replaced by an adjustment equation, such as

$$\frac{dY}{dt} = \beta(M^S - M^D). \quad (2')$$

If β is a constant, then the time pattern of Y will depend solely on initial conditions and the time pattern of M^S . However, if β depends on real variables (including but not limited to interest rates), real factors will again enter in.

¹⁴ Two other comments seem worth making. (a) If in the comment by Johnson numbered [1] above, "real sector" means solely the effect of real factors on interest rates, then the statement is an uninteresting tautology. (b) In my own work, I have generally concluded that the elasticity of real per capita money balances with respect to real per capita income is greater than unity. The essence of my permanent income hypothesis of the demand for money is that equation (2) of the preceding footnote is incorrect if measured income is the income variable in (1), so I have always left still another avenue for real factors to enter.

the conditions given for nominal income if in addition total real output can be regarded as constant.

To turn from the theoretical to the empirical, the evidence I have examined suggests that this kind of divorce is seldom if ever approximated for levels of nominal income or prices over considerable periods. On the other hand, I believe it is frequently approximated for year-to-year and similar changes in nominal income under a wide variety of conditions, and for *changes* in prices under a rather narrower set of conditions.

For the United States, for example, the correlation between year-to-year percentage changes in the nominal quantity of money (defined as currency outside banks plus all commercial bank deposits adjusted) and year-to-year percentage changes in nominal income (defined as net national product) for the ninety-four years from 1870 to 1963 is .70, and the relation seems to have displayed no secular change.¹⁵ The correlations are decidedly lower though still statistically significant for prices—.54 for wholesale prices, and .58 for the price deflator implicit in nominal income.¹⁶ I have seen similar correlations for nominal income for a number of other countries that vary widely in economic structure and financial institutions. The results are almost always very similar.

The closest approximation to a divorce in this sense is observed for conditions of substantial or hyper-inflation. Under such conditions, the

¹⁵ For example, the use of the simple regression corresponding to the above correlation gives the following results for the years from 1962–66 (only data through 1963 were used in computing the regression):

Rate of Change of Nominal Income (Percentage per year, continuously compounded)		
Year	Computed from regression	Actual
1962	5.9	6.9
1963	6.6	5.2
1964	6.4	7.1
1965	7.5	7.8
1966	6.5	8.7

¹⁶ Comparisons for prices like those in the proceeding footnote are

Year	Wholesale Prices		Implicit Price Index	
	Computed	Actual	Computed	Actual
1962	1.6	0.3	1.6	1.1
1963	2.1	−0.3	1.9	1.3
1964	1.9	0.2	1.8	1.7
1965	2.7	2.0	2.3	1.8
1966	2.0	3.2	1.8	3.0

rate of change of prices becomes the most important single variable in the demand function for real money balances, but the rate of change of prices can itself be "explained" by the past history of the quantity of money, so the behavior of both nominal income and prices can be predicted. Changes in real income are likely to be minor compared to changes in the quantity of money.¹⁷

To avoid misunderstanding, let me emphasize the desirability of improving on the approximations just described. Further scientific progress has consisted of and will consist of such improvements. Some improvements have been and will be the development of more sophisticated and accurate methods of allowing for the influence of changes in the quantity of money itself; others have been and will be more precise estimates of the influence of such variables as the real rate of yield on capital, the structure of interest rates, and so on, and the construction of more sophisticated models incorporating explicitly the interrelations between monetary and real magnitudes.

(b) *Does the quantity of money affect real magnitudes?*—The second interpretation of "divorce" is related to the extensive theoretical discussion about the "neutrality" of money that has been so important a feature of the post-Keynesian developments.¹⁸ However, the details of that discussion need not concern us, since the only question at issue here is the effect of including or excluding interest rates in the demand for money on the possibility of regarding money as "neutral."

Consider the IS-LM analysis of the kind first introduced by J. R. Hicks¹⁹ that has become standard in the textbooks. Consider further, the flexible price, full-employment versions.²⁰ In these, interest rates and real per capita income are determined entirely in the real sector; these in turn determine the real quantity of money demanded, which interacts with the nominal quantity supplied to determine the price level. Or to put it another way, interest rates and real per capita income determine velocity; given velocity and real income, changes in the quantity of money can be taken, in Johnson's words, "to affect money income in the short run and prices in the long run without interference from the real forces." Changes in the quantity of money need not affect interest

¹⁷ The already classic study on this point is Cagan, *The Monetary Dynamics of Hyperinflation* in *Studies in the Quantity Theory of Money* 25 (Friedman ed. 1956); since then, other studies by John Deaver for Chile, Adolfo C. Diz for Argentina, Allen Hynes for a number of Latin American countries and Maurice Allais for a large number of countries have confirmed and extended Cagan's results.

¹⁸ For an excellent summary, see Johnson, *Monetary Theory and Policy*, 52 *Am. Econ. Rev.* 334, 343–357 (1962).

¹⁹ Hicks, *Mr. Keynes and the "Classics"; A Suggested Interpretation*, 5 *Econometrica* 147 (1937).

²⁰ A particularly clear exposition is given by Bailey, *National Income and the Price Level* 11–42 (1962).

rates, and so rebound on the real sector, if prices react rapidly enough so that there are no changes in the *real* money stock, which is the desired magnitude affected by interest rates.

To put the matter still more abstractly, the divorce of money from real factors in the sense under discussion requires that there be a way of expressing the equations comprising the theoretical model such that it has a subset of equations sufficient to determine the real magnitudes which do not contain as separate variables either the nominal quantity of money or the price level.²¹ In that case, the system of equations simultaneously determining the real and monetary variables can be dichotomized into one set which determines the level of real income and the interest rate and a second set which together with the solution of the first set determines the level of nominal income and the price level, and this is true *regardless of whether the demand equation for money in the second set has the interest rate as one of its variables*. The “real variables” may then affect the level of nominal income and prices consistent with a given money stock, but the level of the money stock will not affect the real variables, only nominal income and the price level. One of Bailey’s full employment models is a system of precisely this kind.²²

The italicized statement flatly contradicts the comments I have numbered [1] and [4]. Its relation to comment [5] requires a little more exegesis. The first sentence of that comment seems to refer to the sense of divorce discussed in section (a), above. As to the second sentence, I assume that Brill means by “propensities to spend,” saving and investing propensities. In that case, “fluctuations in propensities to spend” will influence equilibrium interest rates, contrary to his assertion, whether or not interest rates enter the demand function for money. These are the real factors affecting interest rates referred to in the preceding paragraph.²³

As to income, Brill’s comment is correct for the usual versions of the IS-LM analysis. On a more sophisticated level, there are ways in which “fluctuations in spending propensities” can influence *real* income for a given money stock whether or not interest rates enter the demand function for money, though even then, if the demand for money is unit-

²¹ The word “separate” is included because it is entirely permissible for $\frac{M}{P}$ or the real quantity of money, to be included, or for M and P to enter as deflators of other nominal magnitudes. To put the condition differently, the requirement is that the reduced form equations of the subset, expressing the real endogenous magnitudes as a function of other variables be homogenous of degree zero in M and F.

²² *Id.* at 35–36.

²³ Perhaps Brill had in mind the strict Keynesian case referred to in footnote 9 in which interest rates are assumed not to affect either investment or saving. In that case, it is true that, if interest rates do not enter the demand function for money, “fluctuations in propensities” will not influence interest rates. However, in that case, interest rates apparently enter nowhere, hence the whole problem has disappeared. In the textual discussion above, I have continued to exclude this strict Keynesian case.

elastic with respect to real income, zero interest elasticity will mean that fluctuations in the propensities do not affect nominal income.²⁴

To avoid misunderstanding, let me emphasize that what I am asserting is that the possibility of constructing a theoretical model in which monetary changes do not impinge on the real sector in no way hinges on whether the interest elasticity of demand for money is zero. I am not asserting that such a model is the most useful, or even *a* useful model to interpret reality. On the contrary, as I should have supposed is abundantly clear from all of my work in the field of money, I am myself persuaded that it is far more useful to introduce interactions between

²⁴ (1) *Full-employment, flexible price version of IS-LM analysis.* In this version, real income is determined entirely by the production sector, including the supply of labor, and is a given number, which is not affected by any monetary change, whether or not the rate of interest enters into the demand function for money. However, this version implicitly assumes that there is only a single commodity, with consumption and capital formation representing simply different uses of the same commodity. While such one-commodity models are very useful for many purposes, they are exceedingly misleading for the present purpose, which is precisely to trace the effects of shifts between consumption and capital formation. If these are different commodities, a shift in propensities means a change in the composition of income. Unless the rate of substitution between the two commodities in production is independent of relative output (in which case, they are perfect substitutes in production and hence can be regarded as a single commodity), there is an index-number problem in determining whether real output has risen or fallen. Clearly, it has been affected in a meaningful way. (I am indebted to Axel Leijonhufvud for this point.) In addition, of course, the change in the rate of accumulation will change real income in every sense in the future.

(2) *Unemployment, inflexible price version of the IS-LM analysis.* This is clearly the more important case for Brill's comment. In the usual text book analysis, the LM curve shows the combinations of interest rates and real income consistent with the given real demand function for money and the given nominal quantity of money (or supply schedule, if interest rates or real income are assumed to affect the nominal quantity of money). If interest rates as well as income enter the demand function, the LM curve has a positive slope. An upward shift in the negatively sloping IS curve produced by a "fluctuation in propensities to spend" will then mean higher interest rates and higher real income; a downward shift, lower interest rates and lower real income; given complete rigidity of prices, nominal income will move in the same direction as real income and by the same percentage. If interest rates do not enter the demand function, and if, as Brill assumes, the nominal quantity of money is fixed, then the LM curve will be vertical, so upward and downward shifts in the IS function affect only the interest rate and not real income or, with rigid prices, money income. This is presumably the analysis underlying Brill's comment and on these assumptions his comment is entirely valid for income though not for interest rates.

However, on a more sophisticated level, there are three channels whereby changes in propensities can influence real income, starting from a position of unemployment. (a) The effects via a shift in composition of output described in (1) above. (b) Price inflexibility need not mean complete rigidity, but rather slow adjustment. The initial unemployment situation presumably means that prices are tending to fall relative to their anticipated behavior. Changes in spending propensities can affect the rate of fall by increasing or decreasing deflationary pressure and this will in turn affect real income. (c) Price inflexibility means that some dynamic adjustment mechanism is at work, which permits actual and desired or temporary and full equilibrium positions to differ for prices. But then, this must be true also of the money market, so changes in the spending propensities can produce (or alter) discrepancies between actual and desired balances, requiring the introduction of an adjustment equation like equation (2') in footnote 13.

the real and monetary sectors than to omit them in the analysis of both long-run growth and short-run cyclical fluctuations. For growth models, the desired capital stock or wealth, and the desired distribution between money and other wealth should, I believe, be made to depend on the rate of change of prices as well as on the real yield of capital; and the desired rate of change in wealth should be made to depend on the discrepancy between actual and desired stocks of wealth (which means that the so-called Pigou effect will operate). For cyclical fluctuations, discrepancies between desired and actual stocks of money (or between anticipated and actual rates of monetary growth) should be regarded as affecting rates of change of both real output and prices. Indeed, I believe they may well be the key element in cyclical fluctuations.

III. Keynes and Interest-Elasticity of Monetary Demand

Johnson's statement, "to admit interest rates into the demand function for money is to accept the Keynesian Revolution and Keynes' attack on the quantity theory," seems to me a misleading interpretation of the history of thought. Keynes' analysis of liquidity preference and of how interest rates affect the quantity of money demanded is certainly a basic contribution to monetary theory and it has stimulated important and valuable research. But this part of his analysis is in the older tradition. Indeed, it was foreshadowed by the strictly quantity theory approach of his *Tract on Monetary Reform*.²⁵ Certainly, Irving Fisher and other classical writers were aware of the effect of interest rates on velocity.²⁶ In my own theoretical essay, *The Quantity Theory of Money—A Restatement*, I emphasize the role of interest rates in the demand function for money without in any way accepting either the Keynesian Revolution or Keynes' attack on the quantity theory.²⁷

The specifically Keynesian innovation in this area, I believe, was the idea that absolute liquidity-preference, that is, the liquidity trap or an infinitely elastic demand for money, might be empirically relevant in deep depressions. As noted earlier, this empirical assertion does have far-reaching theoretical implications. But simply introducing interest rates in the demand function for money does not.

²⁵ Keynes, *Tract on Monetary Reform* 81–95 (1924).

²⁶ For example, Pigou wrote in 1917, "Other things being equal, the variable k ['the proportion of his resources the average man chooses to keep . . . in the form of titles to legal tender'] will be larger the less attractive is the production use and the more attractive is the rival money use of resources. The chief factor upon which the attractiveness of the production use depends is the expected fruitfulness of industrial activity," i.e., the real yield on capital. Pigou, *The Value of Money*, 32 *Q.J. Econ.* 38, 42–46 (1917); reprinted in *Readings in Monetary Theory* 162, 166–168 (1950).

²⁷ *Studies in the Quantity Theory of Money* (Friedman ed. 1956).

IV. Policy Implications

(a) *Fiscal policy*.—Whether, as Brill asserts in the comment I have numbered [6], zero interest elasticity would prevent fiscal policy from affecting “the level of aggregate [nominal] expenditures for goods and services, quite apart from their influence on the stock of money,” depends on the same considerations adduced above in discussing his comment [5]. Fiscal policy could alter interest rates and through them, the composition and level of real income, and these in turn might affect velocity and so aggregate nominal expenditures.²⁸ Whether or not they affected aggregate nominal expenditures, they would affect prices.

(b) The degree of financial restraint, referred to by Brill in the comment I have numbered [7], is not reflected in any simple way by variations in the nominal money stock whether or not interest rates enter into the demand for money. To illustrate, a 10 per cent per year rate of growth of the quantity of money that comes after a period in which the quantity of money has been rising at the rate of 50 per cent per year implies a very different degree of financial restraint than if it comes after a period in which the monetary growth rate has been zero per cent per year. The former may produce a financial panic; the latter a “sloppy” money market, in Federal Reserve terminology.

Points (a) and (b) both illustrate what seems to me the most serious and widespread defect in current discussions of monetary theory and policy—the tendency to neglect the price level and its determination. This is the legacy of Keynes, surely unintended, that has been most productive of misunderstanding.

(c) *Discretion vs. rule*.—There are a set of conditions under which interest rate effects on the demand for money would, as Johnson asserts in the comment I have numbered [3], make the rule of a steady growth in the quantity of money automatically destabilizing.²⁹ They are first, that total output grow at a steady rate without fluctuations, second, that real fluctuations take the form of changes in the ratio of saving and investment to full employment income mediated through interest rate changes, and third, that velocity be a function solely of interest rates. In that case, a rise in interest rates would raise velocity. Since income is growing at a steady rate, the rate of growth of the real quantity of money would have to decline to equate actual with desired balances. Since the nominal quantity of money is growing at a steady rate, a decline in the rate of growth of real balances would in turn require an increased rate of rise of prices. Conversely, whenever interest rates fell, the rate of rise of prices would slow down. Hence fluctuations in interest rates would produce fluctuations in the rate of change of prices. These could

²⁸ See footnote 24.

²⁹ I am indebted to Harry Johnson for spelling out the conditions he had in mind.

be counteracted only by changes in the rate of growth of the quantity of money sufficient to offset the changes in desired velocity.

These conditions are highly special. Moreover, they differ in a number of crucial respects from those that I have assumed to hold in recommending the rule that the quantity of money be made to grow at a constant rate.

(i) Though stressing the importance of monetary change, I have always emphasized “the existence of other factors that affect the course of business or that account for the quasi-rhythmical character of business fluctuations.”³⁰ For the moment, keep the assumption that velocity is independent of all variables except, possibly, interest rates but suppose that “real factors” produce fluctuations (or a tendency to fluctuations) not only in the ratio of investment to output but also in output itself, involving a tendency for interest rates and output to move in the same direction (which roughly corresponds to actual experience). In that case, if velocity were independent of interest rates as well (that is, if velocity were a numerical constant), a constant rate of monetary growth would mean that nominal income would rise at a constant rate, which in turn would mean that prices would decline relative to their trend when output rose and rise relative to trend when output fell—movements that most would regard as automatically stabilizing. Let velocity now be sensitive to interest rates. Movements in velocity would then tend to produce movements in nominal income in the same direction as in output. The amplitude of these movements and hence the direction of movement in prices depends on the amplitude of the interest rate movements and the interest-elasticity of velocity. Prices might still move counter to output, though by less than for a zero interest-elasticity, or they might be stable, or they might move in the same direction. Only in this final case, when prices and output move in the same direction does it seem to me meaningful to speak of the result as in any sense “automatically destabilizing.” In the other cases, the most one can say is that velocity movements offset to some extent the automatically stabilizing effects of the steady rate of monetary growth.

(ii) In practice, velocity does tend to move in the same direction as output, which reflects partly interest rate variations and partly other factors. Also, prices and output tend to move together. However, in practice, the quantity of money also behaves pro-cyclically and so reinforces the movements in velocity. Hence, whether or not the rule

³⁰ Quotation from Friedman and Schwartz, *Money and Business Cycles*, 45 *Rev. Econ. & Stat. Supplement* (February) 55 (1963). See also, Friedman, *The Supply of Money and Changes in Prices and Output*, in *Joint Economic Committee, 85th Cong., 2d Sess., The Relationship of Prices to Economic Stability and Growth* (Compendium of papers submitted by panelists appearing before the Joint Economic Committee) 251 (Comm. Print 1958); Friedman, *A Program for Monetary Stability* 98–99 (1959).

might be “destabilizing” relative to some utopian norm, it has always seemed to me likely to be stabilizing relative to actual discretionary policy.

(iii) Finally, the major argument for the rule has always seemed to me to be far less that it would moderate minor cyclical fluctuations than that it would render impossible the major mistakes in monetary policy that have from time to time had such devastating effects. This consideration has nothing to do with the interest elasticity of demand.

V. Conclusion

It is important that we try to determine as accurately as possible the characteristics of the demand function for money, including the elasticity of demand with respect to interest rates. But in my opinion no “fundamental issues” in either monetary theory or monetary policy hinge on whether the estimated elasticity can for most purposes be approximated by zero or is better approximated by -0.1 or -0.5 or -2.0 , provided it is seldom capable of being approximated by $-\infty$.

The important consideration for monetary theory and policy is whether the demand for money can be treated as a reasonably stable function of a fairly small number of variables and whether this function can be empirically specified with reasonable accuracy. Whether one important argument of the function is an interest rate or set of interest rates is much less important.

The significance that has been attached to the interest elasticity of demand for money reflects, I believe, sophisticated versions of the errors of confusing a movement along a demand or supply function with a shift of such a function and of confusing real with nominal magnitudes. If the interest elasticity is not zero, there will be movements along the function that it is easy to interpret as a sign of instability of the function.³¹ The tendency to treat prices as if they were determined outside the monetary system, or as if they were constant, which may be illuminating for some problems, tends to lead to the neglect of factors that may affect real but not nominal magnitudes.

³¹ Though in practice, perhaps a more important source of confusion of this type is failure to allow for leads and lags.

Government Revenue from Inflation

It has become common to regard inflation produced by the issue of fiat money as a tax on cash balances. The simplest case arises when the government is the sole issuer of money and all money is non-interest bearing.¹ In this case, the real yield from the tax in equilibrium, when holders of cash are fully adjusted to the inflation, is typically taken to be equal to the rate of price rise times the real stock of money, which product in turn is taken as equal to the real value of the new money issued (see Friedman 1953; Bailey 1956; Cagan 1956; Johnson 1967). The rate of price rise is the rate of the tax. The real stock of money is the base of the tax. By strict analogy with an excise tax on a commodity, the yield is the product of the two. As for an excise tax, the yield is a maximum at the rate of tax (that is, rate of inflation) at which the elasticity of demand for real money balances with respect to the rate of inflation is equal to unity. For still higher rates of tax, the tax base declines by a larger percentage than the tax rate rises, so that the product declines.

This analysis is entirely correct for a stationary economy with fixed real income. But it is seriously misleading for a growing economy. For such an economy, the issuer of money obtains a yield from two sources: a tax on existing real cash balances; and provision of the additional real

The key idea of this paper originated in discussions at the Bank of Korea on the appropriate monetary policy for Korea. I am indebted to Governor Jin Soo Suh of the Bank of Korea and other members of the bank staff for hospitality that made my visit, though brief, highly productive. I am indebted also to members of the Money and Banking Workshop of the University of Chicago for helpful comments on an early draft of this paper.

¹ This involves no essential loss of generality. If there is competition in the creation of deposits and no restraints on the payment of interest on deposits, the more complex case involving deposits reduces to this one with "high-powered money" replacing the fiat money of our model. If there is an effective restriction of interest payments on deposits, a broader total than high-powered money replaces the fiat money of our model. In this latter case, the tax yield is shared between the issuers of high-powered money and the issuers (or favored borrowers) of other components of the broader total.

cash balances that are demanded as income rises. As in the stationary economy, the yield from the tax part is maximized at the unit-elasticity rate of price rise. However, the yield from the second part declines monotonically as the rate of price rise increases. As a result, the rate of price rise that will give maximum total yield is lower, and may be drastically lower, for a growing economy than the unit-elasticity rate.²

The question of the rate of inflation that will yield maximum revenue is especially topical for underdeveloped countries embarked on programs of development. For such countries, it has often been argued that inflation is desirable or inevitable. This argument has many facets. The one that is relevant here is that inflation is a form of taxation, the proceeds from which can be devoted to investment. Mundell (1965) and Marty (1967) have demonstrated that the possibilities along this line are extremely limited if there are no other sources of growth. However, when other sources of growth exist, their argument is in one sense too favorable to inflation, in another sense too adverse. It is too favorable because the revenue-maximizing rates of inflation are much lower than they implicitly assume. It is too adverse because growth increases the revenue that can be obtained by the issuance of money.

Of course, the revenue potentiality of money issue is only one element relevant to the desirability of using this device to raise revenue or of using the revenue so obtained to finance investment. This note is devoted solely to the technical issue of the revenue from inflation, not to the broader issue.³

Formal Analysis

Let

M = nominal quantity of money,
 P = price level,
 N = population,

² This point is implicit in Mundell (1965), but he does not recognize its importance because the only growth he considers is the growth which is produced by investment financed by government from the proceeds of the money issue. He assumes all other sources of growth to be zero. Marty (1967) makes explicit what Mundell leaves implicit and even gives the correct formula for total government revenue, but again he fails to recognize the importance of allowing for growth because he follows Mundell in considering only that growth produced by investment financed by the money issue. Both Mundell and Marty conclude that such growth is at best very small.

³ My own personal view is that inflation is neither desirable nor necessary, that the most effective road to development is through free enterprise and private investment, and that the government can serve best by limiting itself to essential government functions, keeping taxes of all kinds low, refraining from intervention into the economy, and providing a stable monetary framework.

$$\begin{aligned}
 m &= \frac{M}{PN} = \text{real quantity of money per capita,} \\
 Y &= \text{nominal income,} \\
 y &= \frac{Y}{PN} = \text{real income per capita.}
 \end{aligned}$$

Let g with a subscript designate a *logarithmic* time derivative of the variable in the subscript, so $g_m = d(\log m)/dt = (1/m)(dm/dt)$ = percentage rate of growth of real quantity of money per capita.

Since we are interested in the equilibrium position after full adjustment to the inflation produced by money issue, we shall regard g_p as simultaneously the actual and the anticipated rate of change of prices.

Let the demand for real money balances per capita be given by

$$m^D = f(y, g_p). \quad (1)$$

Interest rates are omitted for simplicity, though insofar as the real interest rate is affected by the rate of inflation it should be taken into account, since different rates of inflation will then also affect the total yield by altering interest rates and thereby the amount of money demanded. However, this effect seems clearly of a smaller order than those that we shall consider. In the same spirit, we shall neglect any effect of the rate of inflation on the level or rate of growth of real income.⁴ Equation (1) also assumes that the demand for money is homogeneous of first degree with respect to prices and population, so that the demand for aggregate nominal money is

$$M^D = NP \cdot f(y, g_p). \quad (2)$$

Take logs of both sides, differentiate with respect to time, and assume equilibrium, so $M^D = M^S$ (money supplied) = M . This gives

$$g_M = g_N + g_P + \eta_{my}g_y, \quad (3)$$

where η_{my} = elasticity of real per capita money balances with respect to real per capita income. In obtaining equation (3), the time derivative of g_p is treated as zero, which means that equation (3) is valid only for alternative steady states of inflation. The yield from money issue in real terms, call this R (for revenue), is

$$\begin{aligned}
 R &= \frac{1}{P} \frac{dM}{dt} = \frac{M}{P} \cdot g_M = \frac{M}{P} (g_N + g_P + \eta_{my}g_y) \\
 &= N \cdot f(y, g_p) (g_N + g_P + \eta_{my}g_y). \quad (4)
 \end{aligned}$$

⁴ That is, we shall neglect the only effect Mundell and Marty consider.

Note that in the stationary case in which $g_N = g_y = 0$, equation (4) reduces to

$$R(g_N = g_y = 0) = \frac{M}{P} \cdot g_P, \quad (5)$$

or to the usual formulation, with g_P the tax rate and M/P the base.

To determine the value of g_P which maximizes the revenue, differentiate K as given by equation (4) with respect to g_P and set equal to zero:

$$\begin{aligned} \frac{dR}{dg_P} &= N \cdot f(y, g_P) \cdot \left(1 + g_y \frac{d\eta_{my}}{dg_P}\right) + N(g_N + g_P + \eta_{my}g_y) \frac{df(y, g_P)}{dg_P} \\ &= \frac{M}{P} \left(1 + g_y \frac{d\eta_{my}}{dg_P}\right) + \frac{M}{P} \cdot \frac{1}{f(y, g_P)} (g_N + g_P + \eta_{my}g_y) \frac{df(y, g_P)}{dg_P} \quad (6) \\ &= \frac{M}{P} \left[1 + g_y \frac{d\eta_{my}}{dg_P}\right] + (g_N + g_P + \eta_{my}g_y) \frac{d \log m^D}{dg_P} = 0. \end{aligned}$$

The rate of inflation that yields the maximum revenue is then the value of g_P which satisfies the equation,

$$(g_N + g_P + \eta_{my}g_y) \frac{d \log m^D}{dg_P} + g_y \frac{d\eta_{my}}{dg_P} = -1. \quad (7)$$

For a stationary economy, where $g_N = g_y = 0$, this reduces to the usual solution

$$g_P \cdot \frac{d \log m^D}{dg_P} = \eta_{mg_P} = -1, \quad (8)$$

or the point of unit elasticity.

For a growing economy, for which g_N and g_y are both positive, it is clear that equation (7) would not generally be satisfied at the value of g_P for which equation (8) is satisfied. To begin with, assume that the income elasticity is not affected by the rate of inflation, so the final term on the left-hand side of equation (7) is zero. Then at the value of g_P at which equation (8) is satisfied, the first term on the left-hand side would be less than -1 , that is, negative and larger in absolute value than unity. It follows that a lower value of g_P (which would lower the absolute value of the left-hand side) would be required to maximize revenue. Now, let the second term on the left-hand side differ from zero. If it is negative—which is to say if the demand for money with respect to income is less elastic at high than at low rates of inflation—this term will make the whole left-hand side even larger in absolute value and require a still lower g_P . The only case in which the value of g_P satisfying equation (8)

might also satisfy equation (7) is if the income elasticity of money is greater at high than at low rates of inflation. In that case, the positive second term might offset the excess over unity of the absolute value of the first term.

There is no way to solve equation (7) as it stands explicitly for g_p . Similarly, we cannot determine explicitly the rate of monetary growth that will maximize yield. We can solve equation (7) for $g_N + g_p + \eta_{my}g_y$ and substitute the result in equation (3). That gives

$$g_M(\text{for maximum } R) = \frac{-1 - g_y \frac{d\eta_{my}}{dg_p}}{\frac{d \log M^D}{dg_p}}. \quad (9)$$

However, in general, the derivatives on the right-hand side are functions of g_p and must be evaluated at the value of g_p that satisfies equation (7). Hence this expression is not of much use.

A Special Case

To get farther, let us consider the special case in which the demand function for money⁵ is

$$m^D = l(y)e^{-bg_p}. \quad (10)$$

In this special case

$$\frac{d \log m^D}{dg_p} = -b, \quad \frac{d\eta_{my}}{dg_p} = 0, \quad (11)$$

so equation (7) can be solved for g_p :

$$g_p = \frac{1}{b} - g_N - \eta_{my}g_y, \quad (12)$$

and equation (9) becomes

$$g_M(\text{for maximum } R) = \frac{1}{b}. \quad (13)$$

This final result is rather remarkable. For this special case, the rate of monetary growth is determined entirely by the slope of the semi-logarithmic straight line giving the demand function for real money balances for a particular per capita real income. It does not depend at all on the rate of real income growth or the income elasticity of demand. However, these factors determine what rate of inflation will be associated

⁵ This is the functional form used by Cagan (1956) to allow for the effect of g_p . He treats real income as constant, so does not deal with the effect of income.

with the revenue-maximizing rate of monetary growth. The higher the rate of growth of population and of real per capita income, and the higher the income elasticity of demand, the lower the rate of inflation.

Although the revenue-maximizing rate of monetary growth is in this special case independent of the other factors considered, the revenue obtained from monetary growth is not. The revenue is the rate of monetary growth times the real stock of money, or, as a percentage of income, the rate of monetary growth divided by income velocity. The higher the rate of inflation, the higher will tend to be the velocity and hence the lower the revenue. So whatever reduces the rate of inflation associated with the revenue-maximizing rate of monetary growth also raises the revenue.

Illustrative Calculations

The importance of these factors can be illustrated by calculating equation (12) for plausible numerical values of the parameters.

Value of b .—The derivative of $\log m^D$ with respect to the expected rate of change of prices should, in principle, be roughly equal to the derivative of $\log m^D$ with respect to “the” nominal interest rate, since that interest rate, under the equilibrium conditions assumed here, will equal the expected rate of inflation plus the real rate of interest. Accordingly, we can use evidence from studies of the effect on the quantity of money demanded of either differential rates of inflation or different rates of interest.

Elsewhere, I have used an estimate of $b = 10$ as “rather on the high side” (Friedman 1969, p. 42).⁶ This estimate, largely based on computed slopes for interest rates, is over twenty times as large as Cagan’s average estimate of b for the hyperinflations he studies (Cagan 1956, p. 45). However, Bailey has shown that the slope is almost surely much higher at low rates of inflation than at the rates that prevailed during hyperinflations (Bailey 1956, pp. 98–99). More recently, Klein has shown that the slopes generally computed for interest rates are biased downward by failure to allow for interest paid directly or indirectly on demand deposits (Klein 1970, chap. 3).

Klein himself obtains estimates of b that range from 7 to 78, when he makes allowance for both interest paid on demand deposits and the regression bias (Klein 1970, table 10). A collection of studies (Meiselman 1970) yields estimates of the slope with respect to the rate of change of expected prices that vary from 0 to 3, with most in the interval from 1 to 3. These estimates are derived from time series studies for Chile

⁶ Note that b has the dimensions of time so its value depends on the time units in which growth rates are expressed. Throughout, I treat them as per year.

TABLE 1
RATES OF INFLATION YIELDING MAXIMUM REVENUE FOR ALTERNATIVE ASSUMPTIONS

ASSUMPTIONS ABOUT POPULATION AND REAL INCOME GROWTH (% per Year)		REVENUE-MAXIMIZING RATE OF INFLATION (% per Year) If VALUE OF b (Demand Slope) Is					
		2		10		20	
		AND INCOME ELASTICITY IS					
		g_N	g_y	1	2	1	2
0	0	50	50	10	10	5	5
0	5	45	40	5	0	0	-5
0	10	40	30	0	-10	-5	-15
2	0	48	48	8	8	3	3
2	5	43	38	3	-2	-2	-7
2	10	38	28	-2	-12	-7	-17

(Deaver 1970), Argentina (Diz 1970), Korea and Brazil (Campbell 1970), and from a cross-section study for forty-seven countries (Perlman 1970). However, these estimates are probably underestimates of the number we seek for two reasons. First, they are computed for periods when the rate of inflation varied considerably and hence probably do not capture the full reaction to maintained inflation at a given rate. Second, they are all subject to a downward regression bias.

Accordingly, let us use three alternative values of b : 2, 10, and 20. These imply that the revenue-maximizing rates of monetary growth are 50 per cent per year, 10 per cent per year, and 5 per cent per year, respectively. These would also be the rates of inflation if there were no growth.

Income elasticity.—The evidence suggests that the income elasticity is high in countries that are in the stage of rapid economic development and approaches unity for the more developed countries in which the financial structure is highly developed. For the United States from 1870 to 1914, the income elasticity was about 2.0, and similar values have recently prevailed in Japan and Korea. The studies referred to in connection with the value of b generally give estimates of income elasticity between 1 and 2. Accordingly, we may use 1 and 2 as alternative values of the income elasticity.

Population growth.—Alternative values of 0 and 2 per cent per year span much of the range that can be anticipated to prevail.

Real per capita income growth.—Experience here is very variable, from a negligible rate of increase in countries like India, to 2 per cent a year in the United States over long periods, to 10 per cent a year or so over a decade or more in Hong Kong, Taiwan, Japan, and Korea. Let us take three values to illustrate the magnitude of the effect: 0, 5, and 10.

Table 1 gives the revenue-maximizing rates of inflation corresponding

to these assumed values of the parameters. One striking result from this table is the importance of the assumed value of b , especially at low levels. The reason is straightforward. The lower b , the more inelastic the demand for money with respect to inflation at each rate of inflation. The more inelastic the demand, the less the erosion in the base of the tax produced by raising the rate and hence the higher the rate that will yield maximum revenue. As b goes to zero, that rate goes to infinity. I have included so low a value of b as 2 precisely to bring out this point. My impression is that this is well below the value that actually prevails in most developing countries.

A second striking result is that allowance for growth clearly makes a substantial difference. For values of the parameters that are not at all implausible, it turns out that zero inflation or even a negative rate of inflation may yield a developing country more revenue from its monopoly of money issuance than a substantial rate of inflation.

Contrast with Experience

There is no getting away from this conclusion in terms of our analysis. Yet this conclusion seems to run directly counter to casual observation of the behavior of developing countries. How can we explain the observed propensity of developing countries to inflate at rates that seem, on this analysis, higher than the rate that would maximize the government's revenue?

The main explanation, I believe, is shortness of time perspective—the difference between the immediate and longer-run effects of behavior. The preceding analysis is for the long run, for steady states. Start with any steady state, whatever it may be. An increase in the rate of monetary expansion will always increase revenue from the issuance of money for a time. Prices will not react instantaneously to the increased issue; and even when the rate of inflation reacts, real balances will not adjust instantaneously to the new rate of inflation. For a time, therefore, the government's revenue from money creation will be higher on both counts. The part which is a tax on existing real balances will be higher, and so will the amount that people add to their real balances. But this is temporary. As people adjust, they will reduce real balances. The tax revenue may still be higher but it will be offset by lower revenue and perhaps a negative revenue from the adjustment of real balances (velocity will rise or fall less rapidly than earlier). Given time for people to adjust to the new rate of inflation, the end result is certain to be a smaller revenue than was obtained immediately after the beginning of more rapid monetary expansion.

In a sense, the situation is an unstable political equilibrium. Govern-

ments tend to look little farther than the next election. If that election is close, an increase in the rate of monetary expansion is sure to provide the government with more revenue. The negative effects on revenue, let alone on more fundamental economic and social matters, will come later.

That, I believe, is the fundamental explanation why governments so often inflate at a higher rate than the rate that would yield the maximum revenue over a considerable period.

Maximum Revenue and Optimum Monetary Growth

This analysis has so far been concerned entirely with government revenue from money issue. It has not taken into account the welfare loss from inflation—the loss to the holders of money because they are induced to hold less cash than they otherwise would (Bailey 1956). This loss can be interpreted as the cost of collecting the government revenue.

From this point of view, the optimum rate of monetary growth is that which would produce a rate of price decline roughly equal to “the” real interest rate (Friedman 1969, chap. 1). From equation (3) this is given by

$$g_M(\text{optimum quantity of money}) = g_N - \rho + \eta_{my}g_y, \quad (14)$$

where ρ is “the” real interest rate. Comparison with equation (13) shows that, at least for the special case for which equation (13) is valid, the revenue-maximizing and the optimum monetary growth rates are completely independent: the first depends only on b ; the second does not depend on b at all. The two could, by accident, coincide; or either could be larger. However, any revenue-maximizing monetary growth rate that produces inflation is of course larger than the optimum growth rate.

One special case of equation (14) may be of some interest. Suppose that $\rho = g_y$, which there is some basis in experience and theory. In that case,

$$g_M(\text{optimum quantity of money}) = g_N + (\eta_{my} - 1)g_y. \quad (15)$$

Prices decline at a rate equal to the rise in per capita output, which means that factor prices are constant. If, in addition, $\eta_{my} = 1$, the required monetary growth rate is equal to the rate of population growth. Alternatively, if $\eta_{my} = 2$, which is more reasonable for countries in an early stage of development, the required monetary growth rate is equal to $g_N + g_y$, or to the rate of growth of real output. Both of these are monetary rules that have frequently been suggested on very different grounds.

Conclusion

It has always seemed natural to assimilate a government monopoly of fiat money issue with a private monopoly of a product produced at zero cost—Cournot's mineral spring. Yet we have just demonstrated that this is not correct. The owner of the spring will maximize his proceeds by charging a price at which the price elasticity of demand is unity. And this will be true whether the economy is stationary or growing. If it is growing, the only effect will be that his income will also grow, but each year he will maximize his revenue by pricing at the unit-elastic point on his demand curve. Why is this not also true for the government which can issue money at a zero marginal cost? Why will it not also maximize its revenue by pricing at the unit-elastic point on its demand curve, regardless of whether there is growth or not?

The answer is because there are two different prices that are relevant to money issue—one, the goods and services that are given up to acquire a dollar; the other, the number of cents per dollar per year that the money holder must spend to keep his real balances constant. Only the latter is analogous to the price of the mineral water. If the water were priced at zero, the proprietor of the spring would get no revenue regardless of growth. On the other hand, if the annual direct (not alternative) cost of holding money were zero, that is, prices were constant, the issuer of money would still get revenue if there were growth, because he would be able to acquire goods and services with the additional pieces of paper he prints.

The closest private analogy I have been able to construct is with a monopolistic producer and servicer of computers or other durable equipment who sells the equipment but charges an annual fee for servicing it. The higher the service fee, the lower the demand for the equipment at any sales price of equipment. It would pay the monopolist to charge a fee lower than that which maximized revenue from servicing a fixed number of machines in order to raise the number that could be sold each year at a given price.⁷ But even this analogy is not complete because the monopolist has two prices to juggle: his service charge affects the demand for new machines but he can determine the price at which he will sell them. In the money case, the rate of inflation simultaneously determines the total sales price, as it were, of the new equipment that is added to the stock, and also the service charge. The money monopolist has only one variable (the rate of monetary growth) to juggle.

⁷ In general, it will pay him to charge for service at a price equal to marginal cost, since that will maximize the value of the machine to the user, and to obtain his monopoly return entirely from the price he charges for the machine. The "in general" is necessary because of the possibility of using the two-price system as a vehicle for price discrimination.

The distinction between the money monopolist and the mineral spring monopolist is perhaps the point of chief theoretical interest in this note. The point of chief practical importance is the strong presumption that money monopolists have not only imposed welfare loss on their communities by their actual inflationary policy but have obtained less revenue from their monopoly than they would have obtained by less inflation.

References

- Bailey, Martin J. "The Welfare Cost of Inflationary Finance." *J.P.E.* 64 (April 1956): 93–110.
- Cagan, Phillip. "The Monetary Dynamics of Hyper-Inflation." In *Studies in the Quantity Theory of Money*, edited by Milton Friedman. Chicago: Univ. Chicago Press, 1956.
- Campbell, Colin D. "The Velocity of Money and the Rate of Inflation: Recent Experiences in South Korea and Brazil." In *Varieties of Monetary Experiences*, edited by David Meiselman. Chicago: Univ. Chicago Press, 1970.
- Deaver, John V. "The Chilean Inflation and the Demand for Money." In *Varieties of Monetary Experiences*, edited by David Meiselman. Chicago: Univ. Chicago Press, 1970.
- Diz, Adolfo Cesar. "Money and Prices in Argentina, 1953–62." In *Varieties of Monetary Experiences*, edited by David Meiselman. Chicago: Univ. Chicago Press, 1970.
- Friedman, Milton. "Discussion of the Inflationary Gap." In *Essays in Positive Economics*. Chicago: Univ. Chicago Press, 1953.
- . *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine, 1969.
- Johnson, Harry G. "A Survey of Theories of Inflation." In *Essays in Monetary Economics*. London: Allen & Unwin, 1967.
- Klein, Benjamin. "The Payment of Interest on Commercial Bank Deposits and the Price of Money: A Study of the Demand for Money." Ph.D. thesis, Univ. Chicago, 1970.
- Marty, Alvin L. "Growth and the Welfare Cost of Inflationary Finance." *J.P.E.* 75 (February 1967): 71–76.
- Meiselman, David, ed. *Varieties of Monetary Experience*. Chicago: Univ. Chicago Press, 1970.
- Mundell, Robert A. "Growth, Stability, and Inflationary Finance." *J.P.E.* 73 (April 1965): 97–109.
- Perlman, Morris. "International Differences in Liquid Assets Portfolios." In *Varieties of Monetary Experiences*, edited by David Meiselman. Chicago: Univ. Chicago Press, 1970.

The Crime of 1873

I am persuaded history will write it [the act of 1873] down as the greatest legislative crime and the most stupendous conspiracy against the welfare of the people of the United States and of Europe which this or any other age has witnessed. [SENATOR JOHN H. REAGAN 1890]

[The demonetization of silver] was the crime of the nineteenth century. [SENATOR WILLIAM M. STEWART 1889]

In 1873 we find a simple legal recognition of that [the demonetization of silver] which had been the immediate result of the act of 1853. [JAMES LAURENCE LAUGHLIN 1886]

You shall not press down upon the brow of labor this crown of thorns. You shall not crucify mankind upon a cross of gold. [WILLIAM JENNINGS BRYAN 1896]

The Act of 1873 was a piece of good fortune, which saved our financial credit and protected the honor of the State. It is a work of legislation for which we can not now be too thankful. [JAMES LAURENCE LAUGHLIN 1886]

The Coinage Act of 1873, to which these quotations refer, was passed by a vote of 110 to 13 in the House and 36 to 14 in the Senate after lengthy, though superficial, committee hearings and floor debate. It attracted little attention at the time even from members of Congress (including Senator Stewart) who voted for it yet who later attacked it

I am indebted for helpful comments on earlier drafts to Michael D. Bordo, Conrad Braun, Phillip Cagan, Joe Cobb, Harold Hough, David Laidler, Hugh Rockoff, and, as always and especially, to Anna J. Schwartz. In addition, David D. Friedman made a number of helpful suggestions and Rose D. Friedman critically reviewed and helped revise the penultimate draft. Finally, an anonymous referee made a number of very helpful suggestions for revision.

in vitriolic terms as a “grave wrong,” a “conspiracy” perpetrated by “corrupt bargains,” a “blunder which . . . is worse than a crime,” a “great legislative fraud,” and, finally, “the crime of 1873” (see Barnett 1964, pp. 178–81).¹

How did this apparently innocuous legislative measure evoke such strong and contrasting reactions from leading scholars, businessmen, and politicians over so long a period? How did it become a central issue in a presidential campaign conducted more than two decades after its passage? Was it a crime, in any sense of the term? What were its actual consequences? To answer these questions requires some background in monetary history and theory.

I. The Background

The U.S. Constitution gives Congress the power “to coin money, regulate the value thereof, and of foreign coin” and prohibited the states from making “anything but gold and silver coin a tender in payment of debts.” In initially exercising this power, the Congress, following the recommendation of Alexander Hamilton, passed the Coinage Act of April 2, 1792. That act defined the basic monetary unit of the United States as the dollar and defined subsidiary coinage on a decimal basis (the cent, “half disme” [later, the nickel], the “disme” [later, the dime], quarter, etc.). It further defined the dollar as equal to 371.25 grains of pure silver and the \$10 eagle as equal to 24.70 grains of pure gold, authorized *free coinage* of both silver and gold at the specified ratio of 15 to 1, and specified the fraction of alloy to be combined with pure metal in striking the coins.²

I have italicized two terms that are critical to understanding the “crime of ’73.” “Free coinage” is critical because it gave practical content to a specie standard by providing that the government mint would convert specie that individuals chose to bring to the mint into legal tender

¹ According to O’Leary (1960, p. 390), “The first person to use the word ‘crime’ was George M. Weston, the secretary of the United States Monetary Commission of 1876 . . . in his special report, attached to the full report of the commission” published in 1877. Barnett (1964, p. 180) attributes the first use of the full phrase “the crime of 1873” to Senator Henry M. Teller of Colorado on July 10, 1890.

² The act of 1792 stated that “bullion so brought [to be coined at the legal rates] shall be assayed and coined as speedily as may be after the receipt thereof, and that free of expense to the person or persons by whom the same shall have been brought” (Jastram 1981, p. 63). Hence, coinage was “free” in a dual sense: open to all in unlimited amount and without charge. The provision that no charge should be made for coinage is exceptional. Typically, a small charge, called “seigniorage,” is made for the cost of coining. However, the so-called seigniorage charge has sometimes been manipulated and used for purposes other than repaying the cost of coinage, e.g., by ancient “seigniors” (lords) for revenue or by President Franklin Delano Roosevelt as a device for pegging the price of silver.

currency denominated in “dollars” (initially solely in the form of coins; later in paper certificates as well) at the stated metallic equivalent. “Both” is critical because it effectively established the United States on a bi-metallic standard, that is, a monetary standard that authorized free coinage, and hence the use as money, of either of two metals, silver or gold. These two provisions were equivalent to saying that the government would buy all silver or gold offered to it at prices of \$1.2929... per troy ounce of pure silver and \$19.3939... per troy ounce of fine gold, that is, 15 times as much for an ounce of gold as for an ounce of silver; whence the “ratio of 15 to 1.”³

Though either silver or gold could legally be used as money, in practice only silver was so used until 1834. The reason was simple: There was and is a market for silver and gold outside the U.S. Mint: for jewelry, industrial use, coinage by other countries, and so forth. In 1792, the ratio of the market price of gold to the price of silver was almost exactly 15 to 1, the ratio Hamilton recommended. But shortly thereafter, the world price ratio went above 15 to 1 and stayed there (see Jastram 1981, pp. 63–69). As a result, anyone who had gold and wanted to convert it to money would do better by exchanging the gold for silver at the market ratio and taking the silver to the mint than by taking the gold directly to the mint.

To put the matter in another way, if the mint were a two-way street at a 15 to 1 ratio, an obvious get-rich scheme would be to bring 15 ounces of silver to the mint, get 1 ounce of gold, sell the ounce of gold on the market and with the proceeds buy more than 15 ounces of silver, pocket the profit, and keep going. Clearly the mint would soon be overflowing with silver and out of gold. That is why the mint’s commitment under a bimetallic standard is solely to buy silver or gold (i.e., coin freely), though it may, at its discretion, also sell (redeem) either or both metals. The end result was that the United States was effectively on a silver standard from 1792 to 1834. Gold was not used for money at par value, only at a premium. It was too valuable for that purpose. Gresham’s law was in full operation: cheap money drove out dear money.⁴

In 1834, new coinage legislation was introduced in recognition of the changed world market gold-silver price ratio, which by then was about 15.625 to 1. This ratio was repeatedly recommended by the Select Committee on Coins of the House of Representatives from 1832 to 1834,

³ The continuing decimals (.2929..., .3939...) arise because an ounce troy equals 480 grains. Given that a dollar was defined as equivalent to 371.25 grains of pure silver or 24.75 grains of fine gold, 1 ounce of silver was worth 480 divided by 371.25, or \$1.2929..., and 1 ounce of gold, 480 divided by 24.75, or \$19.3939....

⁴ For precision, the “law” must be stated far more specifically, as Rolnick and Weber (1986) point out.

supposedly in the desire “to do something for gold,” which had not long before been discovered in Virginia, North Carolina, South Carolina, and Georgia and “had become of genuine importance to the four southern states” (O’Leary 1937, p. 83). However, rather suddenly, the select committee changed its recommendation to a ratio of 16 to 1, not to do something *for* gold—though it certainly did—but to do something *against* Nicholas Biddle’s Bank of the United States.⁵ This was at the height of the famous “bank war” between President Andrew Jackson and Biddle, which finally resulted in the failure of the bank to obtain a new charter when its original federal charter expired in 1836. As O’Leary put it, the ratio of 16 to 1 was “a golden club. . . used by Jackson and his supporters to belabor their hated enemy, the Bank” (p. 84). The unsatisfactory state of the currency—a mixture of U.S. and foreign silver coins plus paper money issued by state banks, some of doubtful quality—had made the notes issued by the bank a favored medium of exchange. The act of 1834 was expected to weaken the bank by making gold coins an effective substitute for its notes.

Two points are noteworthy about this episode. First, in 1834, “16 to 1” was a golden club; in the 1890s, “16 to 1” was a silver club. Second, in both cases it was wielded by much the same political constituency against much the same political constituency: the largely rural, small-business, lower-class southern and western supporters of Jackson in 1834 and of William Jennings Bryan in 1896 against the bankers, financiers, big business, and urban upper classes of the East and Northeast.

In any event, the adoption of the 16 to 1 ratio—that is, of an official price of \$20.671835... (= 480/23.22) per ounce of fine gold—spelled the end of the reign of silver. From then to the Civil War, little silver was coined. Even subsidiary silver coins became scarce until 1853, when Congress voted to reduce the silver content of subsidiary silver coinage, so that it no longer became worthwhile to melt them down (at least until the Civil War greenback inflation). From 1834 on, gold coins circulated and gold was the effective standard. Despite the increased demand for gold for monetary use, the gold-silver price ratio fell after the Californian and Australian gold discoveries of the 1840s and 1850s so that its status as “cheap money” seemed secure.

The Civil War temporarily ended the reign of gold. The exigencies of war finance led to the introduction of paper money—greenbacks—

⁵ Though the ratio is described as “16 to 1,” that is an approximation. In the 1834 act, the weight of the gold dollar was set at 23.2 grains of pure gold, which gave a gold-silver ratio trivially higher than 16 to 1. The act was amended in 1837 to make the weight equal to 23.22, which gives a ratio trivially below 16 to 1. The reason for the change was to make the percentage of alloy in the minted coin equal to precisely 10 percent. A good source for the early coinage laws of the United States is National Executive Silver Committee (1890). See also U.S. Commission on the Role of Gold (1982, vol. 1, chap. 2).

issued without gold or silver backing and without any promise to redeem them in either metal. Paper, as it were, became the cheap money. Gold, however, continued to circulate, particularly on the West Coast, but of course not on a one-to-one ratio with greenbacks. There was a free market in which the “greenback price of gold” rose above the official legal price, indeed, at the extreme, to more than double the official price. The government required customs duties and some other obligations to be paid in gold; banks provided separate gold and greenback deposits for their clients. In short, gold and greenbacks circulated side by side at a floating exchange rate determined in the market, though greenbacks were clearly the dominant currency for most purposes and in most areas.

At long last, we return to 1873. Movement was afoot to end the greenback episode and “resume” a specie standard. It was time to start tidying up the coinage legislation. The resulting Coinage Act of 1873 listed the coins to be minted. The list included gold coins and subsidiary silver coins but omitted the historical standard silver dollar of 371.25 troy grains of pure silver. Further tidying occurred in 1874.⁶ That was followed by the Resumption Act of 1875 and successful resumption on the basis of gold on January 1, 1879.⁷

The events culminating in resumption in 1879 precisely parallel a corresponding sequence in Britain six decades earlier: a bimetallic standard before 1797 followed by the adoption of an inconvertible paper standard, the demonetization of silver in 1816, and resumption in 1819 on a gold basis (without the 1816 legislation, resumption would have been on silver). The parallelism is not pure coincidence. The initial step—ending convertibility and adopting a paper standard—was a reaction in both countries to the financial pressures of war.⁸ As in the United States, Britain’s decision to return to a specie standard reflected the desire to have a “sound money” and the outrage of the financial community, holders of government bonds, and some economists at the

⁶ The 1873 act included provision for coining a heavier silver trade dollar to be used in trade with Mexico and the Far East, which were on the silver standard. The trade dollar had legal tender status. This was removed in June of 1874 when Congress passed the revised statutes, which provided that no silver coin was to be legal tender beyond the amount of \$5.00 and that any foreign coin was prohibited from being a tender (see Barnett 1964, p. 178). According to Nugent (1968, pp. 98, 134), the coinage legislation was first introduced by Senator John Sherman in 1868. The bill actually passed was initially drafted in 1869, though clearly there were some subsequent changes, and first introduced into the Senate in April 1870.

⁷ For a detailed discussion of the greenback period and resumption, see Friedman and Schwartz (1963, chap. 2).

⁸ It was not the only possible reaction, despite the tendency of many historians to regard what happened as if it had to happen. France was under even greater financial pressure than Britain, yet, “through twenty years of war, at times against half Europe, [Napoleon] never once allowed a resort to . . . inconvertible paper money” (Walker 1896, p. 87).

inflation produced by the departure from a specie standard. Though Britain's choice of gold instead of silver for this purpose was something of an accident, it was a major reason why the United States made the same choice roughly 60 years later.⁹

If resumption in the United States had occurred under the pre-Civil War coinage legislation, silver would have become the "cheap metal" whenever the gold-silver ratio rose appreciably above 16 to 1 (which happened by 1875), and producers of silver would have found it advantageous to bring their silver to the mint rather than to sell it on the market. Similarly, owners of gold coins would have found it advantageous to melt their coins down and sell the gold on the market rather than use them as money at their nominal face value.¹⁰

In practice, neither the conversion of specie into currency at the mint nor the melting down of gold or silver coins is costless. Commonly, a small seigniorage charge is made to cover the expenses of the mint, and melting involves similar costs. In addition, interest is lost because of the delays involved in minting, and trading involves costs in selling gold or silver, and conversely. As a result, the tendency to regard the legal ratio as a precise number so that only one metal can circulate at a time is a fallacy. Just as "gold points" permit exchange rates of two gold standard currencies to fluctuate within a range without producing gold shipments, under a bimetallic standard "gold-silver price ratio points" permit the ratio to fluctuate within a range without producing either a premium on one metal or its complete replacement by the other.¹¹

The omission of any mention of the standard silver dollar in the Coinage Act of 1873 ended the legal status of bimetallism in the United

⁹ David Ricardo, one of the most influential proponents of resumption, initially favored silver though not bimetallism ([1816] 1951, p. 63). In subsequent testimony of 1819 before a committee of Parliament, Ricardo shifted to gold because "I have understood that machinery is particularly apposite to the silver mines and may therefore very much conduce to an increased quantity of that metal and an alteration of its value, whilst the same cause is not likely to operate upon the value of gold" (1952, pp. 390-91; see also p. 427)—a judgment that like so many judgments based on the opinion of technical "experts" proved to be very wide of the mark.

¹⁰ Currently, it pays to bring neither gold nor silver to the mint because both have been replaced by a cheaper money, paper. There still are, however, official prices on the books (\$1.2929 for silver and \$42.22 for gold). The gold holdings of the U.S. government are still valued on the books at the official price. Yet no one would dream of using a silver coin stamped \$1 or a gold coin stamped \$20 as money at these nominal values. They are numismatic items valued at about \$8 and \$475, respectively. I am indebted to Conrad J. Braun for the rough estimates of the current market values of the silver and gold coins.

¹¹ That was the situation in France from 1803 to 1873, during the whole of which time both gold and silver circulated despite market ratios that departed from the French legal ratio of 15.5 to 1. At times, silver was tending to displace gold; at other times, gold was tending to replace silver (Walker 1896, chaps. 4, 5, esp. p. 121). Fisher (1911, chap. 7) presents a rigorous analysis of the theory of a bimetallic standard as well as an illuminating discussion of the experience of France.

States. Had that fateful line not been omitted from the act of 1873, resumption in 1879 would almost surely have been on the basis of silver, not gold. Hence, “the crime of ’73” in the eyes of the proponents of silver.

Those events raise two questions: the less important but easier to answer is, Was there a “crime,” in any meaningful sense? The far more important but far harder to answer is, What would have been the consequences of including the fateful line?

II. Was There a “Crime”?

In 1877, “an editorial in *The Nation* . . . read in part as follows: ‘Mr. Ernest Seyd, a designing bullionist and secret agent of foreign bondholders, came to this country from London in 1873, and by corrupt bargains with leading members of Congress and officers of the Government brought about the demonetization of silver.’ It was said that he brought with him \$500,000 to bribe certain members of Congress and the Comptroller of the Currency” (cited in Barnett [1964, p. 178]). If that had been true, there would indeed have been a crime in every sense of the term. But no evidence has ever been offered to indicate that it was true. Indeed, Seyd was anything but a “designing bullionist.” He was a British bimetallist who objected strongly to the demonetization of silver by the United States (Nugent 1968, pp. 153, 166). No allegation of bribery has ever been made—let alone documented—against any individual member of Congress or government official in connection with the passage of the Coinage Act of 1873. The act was discussed at great length both in committee and on the floor of Congress and openly voted for by large majorities, though later critics claim that the key provision to which they objected was barely mentioned and not further discussed on the floor.¹² In the literal dictionary sense of crime—“an act punishable by law, as being forbidden by statute or injurious to the public welfare”—there was no crime.

On the other hand, in what the Oxford English Dictionary calls a “more general” use of the term—“an evil or injurious act; an offence, a sin”—the existence of a crime is a question of opinion. What is not open to question is that the standard silver dollar was omitted from the list of coins to be minted intentionally, in full knowledge of the likely consequences, and in the belief that those consequences were desirable. That is made clear by Henry R. Linderman, the director of the mint at the time of the passage of the act, in a book published not long there-

¹² They even cite their opponent in support: “As Professor Laughlin states . . . : ‘The Senate occupied its time chiefly on questions of seigniorage and abrasion and the House on a question of the salaries of the officials’” (National Executive Silver Committee 1890, p. 22).

after (1877, chap. 9). In a report to the secretary of the Treasury in November 1872, when the Coinage Act was pending in Congress, he wrote that "the fluctuations in the relative value of gold and silver during the last hundred years have not been very great, but several causes are now at work, all tending to an excess of supply over demand for silver, and its consequent depreciation" (p. 48).

On the consequences of the act, he wrote that "the declaration in the Coinage Act of 1873, that the gold dollar was to be thereafter the unit of value, and the omission of the silver dollar from the coins to be struck under the provisions of the Act, placed the United States upon the single gold standard. . . . The weight of opinion in Europe and America was against the practicability of maintaining a double standard on any basis which might be selected, and in favor of a single gold standard" (p. 44). In a later chapter, he wrote that "the advocates of the restoration of the old silver dollar . . . appear to think that an error, if not a wrong, was committed in discontinuing its coinage; and they desire to correct the same without reference to the question, whether it would be possible to maintain concurrent circulation of gold and silver coins after resumption in 1879" (pp. 100–101).

Further, as Nugent documents in great detail, Senator John Sherman, chairman of the Senate Finance Committee, had been determined to demonetize silver from at least 1867 and had arranged to have a bill to that effect drafted at the end of 1869. From then on, Sherman, Linderman, John Jay Knox (deputy comptroller of the currency and then comptroller), and Secretary of the Treasury George Boutwell cooperated to push a coinage bill that included the demonetization of silver (Nugent 1968, pp. 80, 88, 99, 103, 105). "Were Knox, Linderman, Boutwell, Sherman, and others aware of what they were doing when they planned to drop the silver dollar?" asks Nugent. "It is inconceivable," he goes on, "that they were not . . . But did they urge it because they feared a drop in silver prices? No one made an explicit statement to that effect, but it was undoubtedly the case" (p. 137).

In addition, as Francis Walker (1893, pp. 170–71) wrote two decades later: "So completely without observation was this measure passed, that it was not for a year or two that the fact of demonetization was popularly known." In an attached footnote, he added that "the writer was in 1873 Professor of Political Economy at Yale, and was actually engaged in lecturing upon the topic of money. He was, also, a pretty good newspaper reader, and by the accidents of position and personal acquaintance, was fairly well in touch with the men of commerce and banking in the neighboring city of New York. Yet it was long after the passage of the act of 1873 that he first learned of the demonetization of the silver dollar" (p. 171, n. 1).

As O'Leary (1960, p. 392) summarized the evidence,

it seems only reasonable to conclude that the failure to include provision for the standard silver dollar in the Coinage Act of 1873 was based not upon recognition of the existing economic facts but rather upon calculated hostility to silver as a part of the monetary standard. The Act anticipated the future. It was purposive and deliberate in the mind of the man [according to Nugent, "men"] who largely framed the legislation and saw it through the Congress. In this sense, the silver people are correct in holding that it was the result of "malice aforethought." It was expected to accomplish and did accomplish a result going far beyond a mere "tidying-up of our coinage laws and procedures."

For the next twenty-seven years the silver question bedeviled the politics and the finances of the United States. Silver never won back the place it would have enjoyed had the Act of 1873 not failed to include provision for the coinage of the standard silver dollar. The consequences of not striking down the free and unlimited coinage of the silver dollar could have been vast for subsequent American financial, economic, and political life. That is, however, another story.¹³

To which we now turn.

III. The Consequences of the Coinage Act of 1873

Eliminating the free coinage of silver had major consequences because of one central fact cited by Linderman: the likely decline in the world price of silver relative to that of gold. Had there been no decline in the silver-gold price ratio—or, as it is more usually expressed, rise in the gold-silver price ratio—it would have been irrelevant whether the fateful line was included in the act of 1873 or omitted. In either event, the pre-Civil War situation of an effective gold standard would have continued when and if the United States resumed specie payments.

As it was, however, a rise in the gold-silver price ratio started well

¹³ In a fascinating paper, Rockoff (1990) persuasively argues that Frank Baum's *The Wonderful Wizard of Oz* "is not only a child's tale but also a sophisticated commentary on the political and economic debates of the Populist Era" (p. 739), i.e., on the silver agitation generated by the so-called crime of '73. The land of Oz, according to Rockoff, is "the East, [in which] the gold standard reigns supreme and in which an ounce (Oz) of gold has almost mystical significance" (p. 745). He goes on to identify the Wicked Witch of the East with Grover Cleveland, the gold Democrat who, as president, "led the [successful] repeal of the Sherman Silver Purchase Act [of 1893]" (p. 746). Similarly, Rockoff is able to identify many of the other places and characters, and much of the action, with places, people, and events that played a significant role in the final years of the free-silver movement.

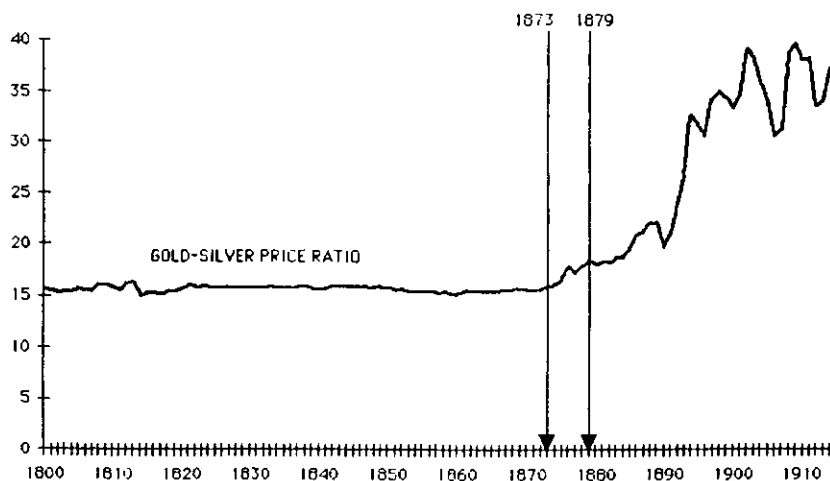


FIG. 1.—Ratio of price of gold to price of silver, annually, 1800–1914.

before the United States passed the act of 1873 and was in full swing when the United States resumed specie payments in 1879. Resumption by the United States on the basis of gold was the final nail in the coffin of silver. The gold-silver price ratio, plotted in figure 1, fluctuated around 15.5 (the mint ratio in France) for decades before the gold discoveries in California in 1848 and in Australia in 1851. It then fell to a low of nearly 15 by 1859, when it started an irregular but more or less steady rise.¹⁴ The rise speeded up rapidly after 1870, as one European country after another shifted from a silver or bimetallic standard to a single gold standard—a tribute to the leadership of Britain, by then the recognized dominant economic power. Germany shifted in 1871–73, after it defeated France and imposed a large war indemnity payable in funds convertible into gold. France, which had maintained a bimetallic standard since 1803, despite first major silver and then major gold discoveries, demonetized silver along with the other members of the Latin Monetary Union (Italy, Belgium, and Switzerland) in 1873–74. The Scandinavian Union (Denmark, Norway, and Sweden), the Netherlands, and Russia followed suit in 1875–76 and Austria in 1879. By the late 1870s, India and China were the only major countries on an effective silver standard. The resulting increased demand for gold and increased supply of silver for nonmonetary purposes produced a dramatic rise in the gold-silver price ratio. From 15.4 in 1870, it jumped

¹⁴ Though France doubtless adopted the ratio of 15.5 to 1 because that was roughly the market ratio in 1803, France's successful maintenance of bimetallism undoubtedly helped to stabilize the ratio (see Walker 1896, p. 87; Fisher 1911, p. 136).

to 16.4 by 1873, 18.4 by 1879, and 30 by 1896, when 16 to 1 was the Bryan battle cry.

By joining the movement to gold, the United States added to the upward pressure on the gold-silver price ratio, both by absorbing gold that would otherwise have been available for monetary use by the rest of the world and by failing to absorb silver. The effects were far from trivial. In preparation for resumption, the U.S. Treasury had accumulated gold so that by 1879, the stock of monetary gold in the United States, both in the Treasury and in private hands, already amounted to nearly 7 percent of the world's stock. By 1889, the U.S. share had risen to nearly 20 percent. Even more dramatically, the increase from 1879 to 1889 in the U.S. stock of monetary gold *exceeded* that in the world's stock. The monetary gold holdings of the rest of the world declined from 1879 to 1883 and then rose but did not surpass its earlier level until 1890.

For silver, the failure to absorb silver via free coinage was offset to some extent by repeated special legislation for the benefit of silver interests requiring the federal government to buy silver at market prices. The first such measure preceded resumption, the Bland-Allison Act of 1878, which authorized the Treasury to buy between \$2 and \$4 million of silver each month at the market price and which led to continued purchases from 1878 to 1890. Then silver purchases were stepped up drastically under the Sherman Silver Purchase Act, until the silver purchase clause was repealed in 1893.

Interestingly enough, the number of ounces of silver purchased under these acts was almost equal to 16 times the number of ounces of fine gold added to the country's monetary gold stock. On first blush, it looks as though political measures absorbed as much silver as free coinage would have. However, that is not the case. As will become apparent in what follows, had the United States been on silver, the stock of money would have risen faster than it did and hence the ounces of silver brought to the mint would have substantially exceeded 16 times the ounces of gold actually acquired.¹⁵

The most obvious, but by no means the most important, consequence of the return of the United States to gold rather than to a bimetallic standard was the sharp rise in the gold-silver price ratio. A far more important consequence was the effect on the nominal prices of goods and services in general. The increased world demand for gold for monetary purposes coincided with a slowing in the rate of increase of the world's stock of gold and a rising output of goods and services. These forces put downward pressure on the price level. Stated differently, with

¹⁵ According to the estimates discussed in Sec. IV below, 26 times as many ounces of monetary silver would have been accumulated as the ounces of gold actually acquired.

gold scarcer relative to output in general, the price of gold in terms of goods went up, which means that the nominal price level (under a gold standard, the price level in terms of gold) went down. The downward pressure was relieved somewhat by a rapid expansion of the banking system, which increased the amount of money that could be pyramided on each ounce of gold. On the other hand, rising real income, plus the spreading monetization of economic activities, plus the declining price level itself increased the downward pressure on prices by leading the public to hold larger cash balances relative to their income (i.e., velocity declined).

The outcome was deflation from 1875 to 1896 at a rate of roughly 1.7 percent per year in the United States and 0.8 percent per year in the United Kingdom, which means in the gold standard world. In the United States, the deflation from 1875 to 1896 followed the even sharper deflation after the Civil War. That sharper deflation was an essential requisite for successful resumption on gold at the prewar parity between the U.S. dollar and the British pound. It also produced wide unrest and dissatisfaction, particularly in rural areas. The unrest led to the formation in 1876 of the Greenback party to continue earlier agitation to issue more greenbacks to replace deflation by inflation. The political agitation ended the retirement of greenbacks, which had started after the Civil War, and led to the adoption in 1878 of the Bland-Allison bill authorizing the Treasury to purchase a limited amount of silver at market prices.

Though the silver was purchased at market prices, it was valued for monetary purposes at the higher legal price, the difference being treated as seigniorage. The silver was mostly coined into standard silver dollars. However, most of the coins were stockpiled in the Treasury as "reserves" for pieces of paper called silver certificates or, after 1890, Treasury notes of 1890. These were nominally convertible into silver, but they were also "legal tender" effectively convertible into gold. Hence, it was cheaper to get silver by using the paper money to buy it on the market than by converting the paper money into silver at the fictional legal price. In effect, the silver certificates were fiat money differing from greenbacks only because the historic role of silver as money made it more acceptable to increase the money supply by buying silver rather than by openly issuing fiat money. It also had the political effect of harnessing silver interests to the populist cause of inflation. The stock of silver in the Treasury was the counterpart to the stock of wheat currently held by the U.S. government as a result of its attempt to prop up the price of wheat.

A 1.7 percent a year decline in prices may seem too mild to generate the kind of agitation that bedeviled the two decades from resumption to the end of the century. But several considerations argue otherwise. First,

the 1.7 percent refers to a price index that covers all goods and services (the implicit price deflator). The wholesale prices of agricultural and other basic commodities doubtless fell at a greater rate (3.0 percent a year by one index). At least as important, we all want the prices of the things we sell to go up, not down, so that sellers of goods and services are almost invariably inflationists. True, we want the prices of the things we buy to go down. But as consumers, we buy many things whose prices are moving in different directions, so we are far less acutely aware of what is happening to the price level than of what is happening to the specific prices of the things we sell. And that was far truer in the nineteenth century, when data on the economy as a whole were few and far between, than it is now. Moreover, at all times, sellers tend to be relatively few in number and to be organized, so that they have more political clout than the dispersed consumers who benefit from declining prices. That was particularly true of producers of silver, who clearly had much to gain by the adoption of a silver standard. Though few, they were politically influential because the sparsely populated silver states had the same representation in the Senate as the densely populated urban states.

An additional factor was that farmers are generally net monetary debtors and, as such, are harmed by a fall in prices, which raises the real value of their debt, and benefited by a rise in prices, which reduces the real value of their debt. As debtors, they were particularly susceptible to propaganda representing the "crime of '73" as the evil machinations of a cabal of eastern and foreign capitalists: Wall Street versus Main Street.¹⁶

One paradoxical result of the agitation for inflation via silver was that it explains why *deflation* was more severe in the United States than in the rest of the gold standard world (1.7 percent vs. 0.8 percent). As Anna Schwartz and I concluded,

This entire silver episode is a fascinating example of how important what people think about money can sometimes be. The fear that silver would produce an inflation sufficient to force the United States off the gold standard made it necessary to have a severe deflation in order to stay on the gold standard. In retrospect, it seems clear that either acceptance of a silver standard at an early stage or an early commitment to gold would have been preferable to the uneasy compromise that was maintained, with the uncertainty about the ultimate outcome and the consequent wide fluctuations to which the currency was subjected. [Friedman and Schwartz 1963, pp. 133–34]

¹⁶ I owe this comment to Hugh Rockoff.

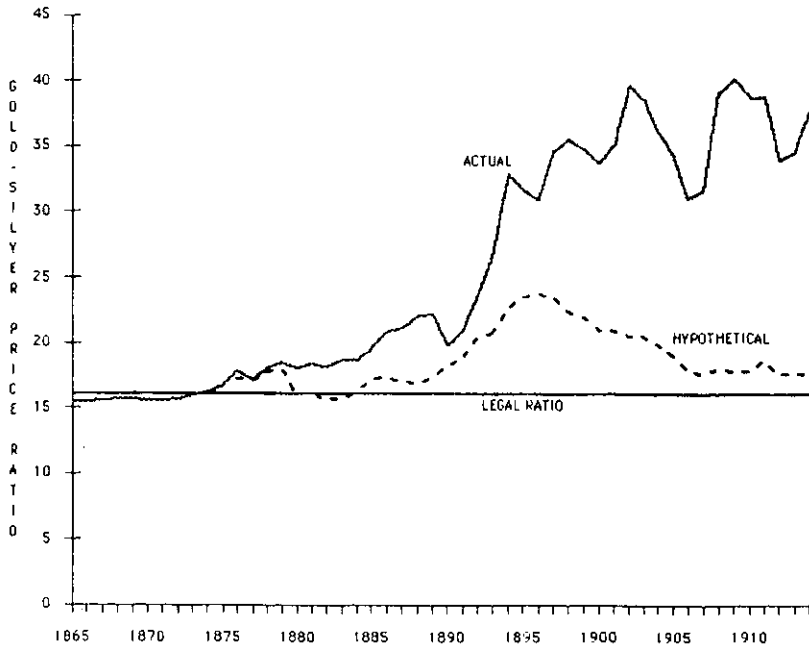


FIG. 2.—Gold-silver price ratio: legal, actual, and hypothetical, 1865–1914.

IV. Which Would Have Been Better: Silver or Gold?

Given that either extreme would have been preferable to the uneasy compromise, which extreme would have been better: the early adoption of silver as the single standard at the monetary value of \$1.2929... an ounce or the early commitment to gold as the single standard? Or, seemingly different from either extreme, the continuation of nominal bimetallism? An answer requires a thorough examination of the quantitative consequences of the three choices.

As it happens, that examination, presented in the Appendix, makes it clear that resumption under a continuation of the bimetallic standard would have occurred to silver, not to gold, and would have occurred in 1876, a year after the passage of the Resumption Act. As a result, the gold-silver price ratio would have behaved very differently than it did.

Figure 2 plots the legal gold-silver price ratio (16 to 1), the actual market price ratio, and an estimate of the hypothetical price ratio that would have prevailed if legal bimetallism had continued. The actual ratio skyrocketed, especially after 1890, when it rose to more than 30 and stayed there. In sharp contrast, the estimated hypothetical ratio departs widely from the legal ratio only from 1891 to 1904. Before 1891, it fluctuates narrowly around 16 to 1. From 1906 to 1913, it remains

between 17 and 18. The years during which the ratio departs widely from 16 to 1 are no accident. The ratio rises well above 16 to 1 during the years of maximum political agitation about free silver surrounding the Bryan free-silver campaign of 1896, and the subsequent unwinding of the effects of that agitation. If the critical line had been retained in the Coinage Act of 1873, that agitation would never have occurred because the United States would have been on a silver standard. The hypothetical price ratio falls to a lower level during the period in which world gold production, which started rising rapidly in 1897, reached peak levels. This tended to depress the real price of gold.

These estimates allow as fully as I could for the alteration in economic circumstances that would have been produced by the continuation of legal bimetallism: the higher world price level and lower real price of gold, the reduction in the amount of silver available for nonmonetary use, and so on. But I have not been able to allow for some predictable effects—notably changes in real income and in the production of silver and gold—let alone for the change in the political climate. No doubt, the political vacuum created by the disappearance of the free-silver issue would have been filled by other issues—very likely pressure for the United States to convert to a gold standard—but there is no way of conjecturing what effect they would have had on the gold-silver ratio. Any attempt to do so would carry this exercise in history as it might have been into the realm of fantasy.

I conclude that the adoption of silver would in practice have produced ratios throughout the period that would have fluctuated not far from 16 to 1 and would have varied even less than the estimates for the period before 1891 and after 1904. In short, I believe that the United States could have played the same role after 1873 in stabilizing the gold-silver price ratio that France did before 1873.¹⁷ If I am right, the fears of the opponents of bimetallism that a bimetallic standard would involve continual shifting between silver and gold would have proved false. With the United States effectively on silver and the United Kingdom and other major countries on gold, changes in the gold-silver ratio would have been directly reflected in the exchange rate between the dollar and other currencies. A rise in the ratio would have produced a dollar depreciation; a decline in the ratio, a dollar appreciation. Here again, a relatively steady gold-silver ratio would have meant relatively steady exchange rates, varying for sterling not far from the level, \$4.86, that actually prevailed.

The gold-silver price ratio is of no great importance in and of itself—

¹⁷ From 1803 to 1873, when France successfully maintained a bimetallic standard at a legal gold-silver ratio of 15.5 to 1, the lowest market ratio was 15.19 in 1859 and the highest 16.25 in 1813; most of the time the range was much narrower (Warren and Pearson 1933, p. 144, table 25).

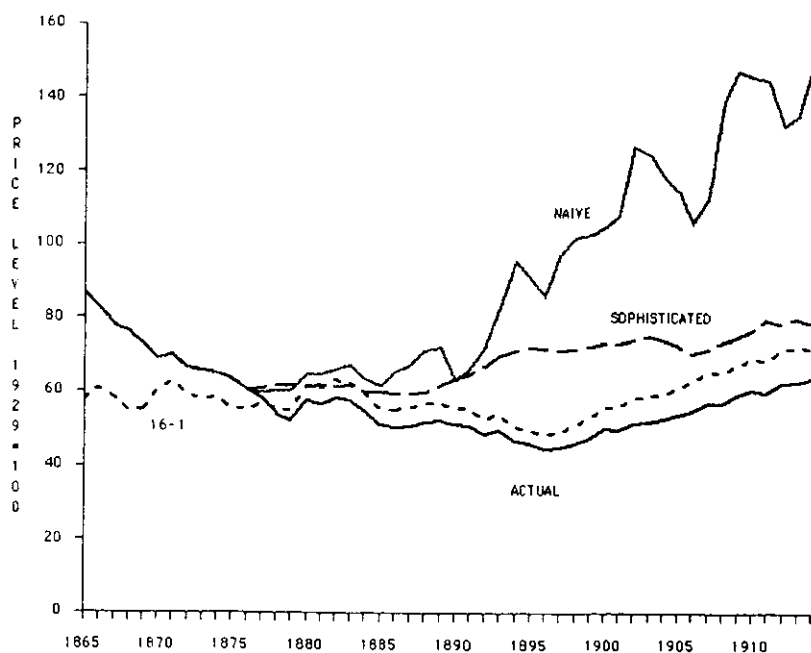


FIG. 3.—U.S. price level: actual and alternative estimates under a silver standard, 1865–1914.

silver and gold dealers excepted—but is of vital importance for the price levels that would have prevailed in silver standard countries (by assumption, including the United States) and in the gold standard countries. Figure 3 plots the actual price level and alternative hypothetical price levels corresponding to the gold-silver price ratios in figure 2. The naive estimate simply assumes that the gold-silver price ratio and the real price of silver would have been what they actually were. On that assumption the price level is readily calculated. It is necessary only to multiply the price level that actually prevailed by the ratio of the legal price of silver (1.2929) to the market price. However, this naive estimate clearly produces a great overestimate of the price rise that would have occurred. The 16 to 1 estimate goes to the other extreme: it underestimates the effect of the adoption of a silver standard on the price level by assuming that the actual ratio would have been precisely 16 to 1 throughout. The “sophisticated” estimate is in between but considerably closer to the 16 to 1 estimate than to the naive estimate for most of the period. However, the 16 to 1 estimate probably gives a more accurate picture of the likely year-to-year pattern than either of the other estimates. Both the naive and the sophisticated estimates are bedeviled by much purely statistical noise. In addition, U.S. bimetallism would have

provided an incentive for worldwide stabilizing speculation in silver that would have eliminated erratic movements.

The U.S. price level fell from 1876 to 1896 at a rate of 1.5 percent a year and rose from then to 1914 at a rate of 2.0 percent a year. The 16 to 1 price level first fell at 0.7 percent to 1896 and then rose at 2.3 percent a year to 1914. The sophisticated price level declined from 1876 to 1887 at a rate of 0.2 percent a year and then rose to 1914 at the rate of 1.1 percent a year. Either alternative implies a halving or more of the initial rate of decline. The 16 to 1 alternative implies a slightly more rapid subsequent rise; the sophisticated alternative a much milder rise. If my estimates are anywhere near correct, a bimetallic—or really silver—standard would have produced a considerably steadier price level than the gold standard that was adopted.

Perhaps even more important, a silver standard almost surely would have avoided what Schwartz and I dubbed “the disturbed years from 1891 to 1897” (Friedman and Schwartz 1963, p. 104): encompassing the very sharp contraction of 1892–94, a brief and mild recovery from 1894 to 1895, followed by another contraction from 1895 to 1896,¹⁸ widespread bank failures plus a banking panic in 1893, and a run on U.S. gold reserves by foreigners fearful that silver agitation would force the United States off the gold standard. Confidence was restored and departure from gold prevented by a private syndicate headed by J. P. Morgan and August Belmont, under contract to the U.S. Treasury. “The allegedly onerous terms of the contract, arranged secretly through agents long identified in Populist literature as ‘the conspiracy of international bankers,’ became an issue in the campaign of 1896” (p. 112 n).

The effects would not of course have been limited to the United States. I have not been able to make as thorough an empirical study for the rest of the world as for the United States, but in the course of preparing the estimates for the United States, it was necessary to estimate the effect on the price level in the gold standard world, for which I used the United Kingdom as a proxy. Figure 4 gives the actual and hypothetical levels of prices in the United Kingdom. The estimated effect, though smaller than that in the United States, is clearly substantial. The price level would have been consistently higher for the rest of the world. The 0.8 percent a year decline in the actual price level from 1875 to 1895 would have been cut to 0.5 percent; the 0.9 percent a year subsequent rise would have been increased to 1.1 percent. Here too, however, there clearly would have been effects other than those encompassed in our simple calculation. The changes in the United States would doubtless have produced echoes elsewhere. Presumably, a healthier U.S. economy would have meant a healthier world economy. In addition, the consis-

¹⁸ These are the annual reference dates used in Friedman and Schwartz (1982).

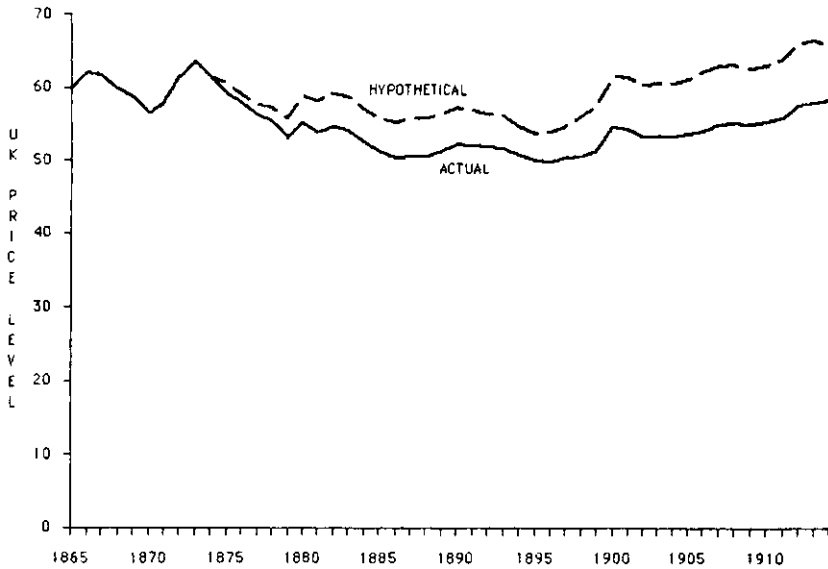


FIG. 4.—U.K. price level (1929 = 100): actual and hypothetical under a U.S. silver standard, 1865–1914.

tently lower real price of gold would have reduced the incentive to produce gold. That might have delayed the introduction of the cyanide process for extracting low-grade ore that was responsible for the flood of gold that produced worldwide inflation after 1896. I have not allowed for any such effect.

Whether or not a verdict of “guilty” would have been appropriate in a court of law for “the crime of ’73,” it is appropriate in the court of history. The omission of the fateful line had momentous consequences for the subsequent monetary history of the United States and, indeed, to some extent, of the world. The rhetoric was overheated, but the importance of the issue was not overstated. The real issue was the monetary standard: gold and silver bimetallism, which in practice in the United States had meant alternating silver and gold standards. The act of 1873 cast the die for a gold standard, which explains its significance. Moreover, while the conventional view is Laughlin’s, that “the act of 1873 was a piece of good fortune” ([1886] 1896, p. 93), my own view is that it was the opposite: a mistake that had highly adverse consequences.

I hasten to add that this is a judgment about 1873, not 1896. By 1896 it was too late to undo the damage. Bryan was trying to close the barn door after the horse had been stolen. The market gold-silver price ratio was about 30 to 1 at the time of the 1896 campaign. Adopting a legal ratio of 16 to 1 would doubtless have produced a sharp drop in the

market ratio by simultaneously reducing the demand for gold and increasing the demand for silver. But however sharp the decline, it would have produced major disturbances in both price levels and international trade. The gold standard world was entering into an inflationary episode; loosening a flood of silver as money could only intensify inflation in the gold standard countries, while releasing it full force on the United States.¹⁹

I hasten to add that this judgment is not intended either to denigrate or to praise the character or the intentions of the various parties in the long-running dispute. The prosilver group contained silver producers seeking to promote their special interests, inflationists eager to seize any vehicle for that purpose, and sincere bimetallists desiring neither inflation nor deflation who were persuaded that bimetallism was more conducive to price stability than monometallism. Similarly, the progold group contained producers of gold; deflationists, pilloried by the free-silver forces as Wall Street bankers; and sincere believers that the gold standard was the only satisfactory pillar for a financially stable society. Motives and intentions matter far less than the outcome. And, in this as in so many other cases, the outcome was very different from that intended by the well-meaning advocates of the Coinage Act of 1873.

Appendix A

Estimating the Effect of Continuing Bimetallism after 1873²⁰

The objective is to estimate the price level and gold-silver price ratio that would have prevailed if the Coinage Act of 1873 had contained provision for free

¹⁹ The estimate in the Appendix is that the market ratio would have been about 24 to 1 in 1896 if the United States had remained on a bimetallic standard. However, as indicated in the text, I suspect that that is a considerable overestimate since it reflects the effect of the monetary and political controversies that would have been avoided under the hypothetical circumstances. Moreover, theoretical considerations suggest that the market ratio that would have prevailed under the hypothetical circumstances might have been either higher or lower than the market ratio that would have been produced by the election of Bryan and adoption of the free coinage of silver and gold at a 16 to 1 ratio. In another paper (Friedman 1990), I have discussed in greater detail what the effect would have been if Bryan had been elected in 1896.

²⁰ Many years ago, I suggested to Louis Drake that he estimate the effect on U.S. and world prices if the United States had remained on a bimetallic standard. He worked on the project for years and accumulated much data, but he was never sufficiently satisfied with his results to publish them. After his death in 1982, colleagues and friends edited a preliminary paper that was found in his files, retaining in full his original calculations, and published the result in *Explorations in Economic History* (Drake 1985). When I began the paper to which this memorandum is appended, I thought that I could simply use his results. But when I read his paper in detail, I appreciated the reservations about his results that presumably led him to refrain from publication. In consequence, I have produced an independent set of estimates, though benefiting from some of his data and analysis. Not surprisingly, my final results differ drastically from his.

coinage of the standard silver dollar of 371.25 troy grains of pure silver, so that the legal market price of silver was 1.2929....

A. *Naive Estimate*

The real price of silver is simply the nominal price divided by the price level. On the naive assumption that it remained unchanged, the real price of silver would have been 1.2929/PHN, where PHN is the naive estimate of the hypothetical price level under a silver standard. Equating the two and solving for the naive hypothetical price level give

$$\text{PHN} = 1.2929 \dots \cdot \frac{P}{\text{PS}}, \quad (\text{A1})$$

where P is the actual price level and PS is the actual nominal price of silver.²¹ The naive estimate is less than the actual price level from 1865 to 1876. In 1876, the two are equal; hence if the fateful line had not been omitted from the Coinage Act of 1873, resumption on the basis of silver would have occurred in 1876, a year after the passage of the Resumption Act. Figure 3 of the text plots the subsequent naive estimate of the price level; table A1 gives the numerical values.

There are three defects of the naive estimate. (1) The United States would probably have added to its silver stock under a silver standard even more than it did in response to the silver interests under a gold standard. That would have tended to raise the real price of silver. (2) The United States would also have exported gold rather than accumulated gold, which would have added to the rest of the world's monetary and nonmonetary stocks of gold and raised nominal prices in the gold standard world. That would have lowered the real price of gold. (3) On both scores, the gold-silver price ratio would have been lower than it actually was.

B. *16 to 1 Estimate*

Assume that the adoption of a silver standard by the United States would have been effective in establishing 16 to 1 as the actual gold-silver price ratio and that the United States stayed on a strict silver standard (i.e., the ratio was trivially above 16 to 1). As we shall see, this is not as farfetched as it seems.

To estimate the hypothetical U.S. price level under this assumption, we need an estimate of the hypothetical real price of gold. Assume that the United States disposed of the whole of its monetary gold stock when it adopted a silver standard and that the gold released was divided between nonmonetary use (by the United States plus the rest of the world) and the monetary gold stock of the rest of the world in the proportion that actually prevailed between these two components

²¹ For sources of data for these and succeeding variables, see App. B. The notation is also defined in App. B.

Year	U.S. PRICE LEVEL				U.K. PRICE LEVEL (1929 = 100)		GOLD-SILVER PRICE RATIO	
	Actual	Naïve	Hypothetical		Actual	Hypothetical	Actual	Hypothetical
			16 to 1	Sophisticated				
1865	86.5		57.3		59.8		15.4	
1866	82.6		60.8		62.0		15.5	
1867	77.6		58.2		61.6		15.6	
1868	76.2		54.6		59.8		15.6	
1869	72.7		54.7		58.7		15.6	
1870	68.7		59.8		56.3		15.6	
1871	69.8		62.5		57.8		15.6	
1872	66.3		59.0		61.4		15.6	
1873	65.5		57.6		63.5		15.9	
1874	64.8		58.3		61.5		16.2	
1875	63.3		55.1		59.2		16.6	
1876	60.4	60.2	55.4	60.4	57.8	59.0	17.8	17.2
1877	58.2	59.8	56.8	60.8	56.2	57.5	17.2	17.1
1878	53.9	60.1	55.4	61.4	55.2	57.2	17.9	17.7
1879	52.0	60.0	54.7	61.4	52.8	55.6	18.4	18.0
1880	57.4	64.5	61.3	61.0	55.0	58.7	18.0	15.9
1881	56.3	64.4	60.8	61.5	53.8	58.1	18.3	16.2
1882	58.1	66.0	62.8	61.1	54.6	59.0	18.2	15.5
1883	57.4	66.8	62.3	61.3	52.5	58.6	18.7	15.7
1884	54.4	63.2	59.1	59.7	52.5	57.0	18.7	16.2
1885	50.8	61.6	55.4	59.6	51.1	55.7	19.4	17.2

1886	50.1	65.1	54.8	59.2	50.3	55.0	20.9	17.3
1887	50.6	66.8	55.8	59.2	50.5	55.7	21.1	17.0
1888	51.5	70.9	56.7	59.7	50.5	55.6	22.0	16.8
1889	51.8	71.5	56.8	61.3	51.2	56.1	22.0	17.3
1890	50.8	62.8	55.8	63.4	52.1	57.1	19.7	18.2
1891	50.3	65.8	55.0	64.4	51.9	56.7	20.9	18.7
1892	48.3	71.3	52.5	66.7	51.8	56.2	23.8	20.3
1893	49.5	81.8	53.8	69.4	51.5	55.9	26.6	20.6
1894	46.4	95.2	50.0	70.9	50.6	54.6	32.9	22.7
1895	45.7	90.5	49.0	71.9	49.9	53.5	31.9	23.4
1896	44.4	85.5	48.1	71.3	49.7	53.9	30.5	23.7
1897	44.6	96.5	48.6	71.1	50.2	54.6	34.5	23.4
1898	45.9	101.8	51.0	71.3	50.5	56.1	35.5	22.4
1899	47.1	102.3	52.6	72.0	51.2	57.1	34.8	21.9
1900	49.6	104.5	55.8	72.6	54.6	61.4	33.8	20.8
1901	49.3	108.1	55.7	72.8	54.2	61.2	35.1	20.9
1902	51.0	126.5	57.9	74.0	53.3	60.4	39.7	20.5
1903	51.5	124.4	58.6	74.9	53.2	60.4	38.6	20.5
1904	52.3	118.2	59.4	73.8	53.3	60.5	36.2	19.9
1905	53.4	114.4	60.8	72.4	53.6	61.0	34.3	19.0
1906	54.5	105.4	62.7	70.1	54.0	62.2	31.0	17.9
1907	56.8	112.4	65.2	71.3	54.9	63.0	31.8	17.5
1908	56.7	138.6	67.0	73.1	55.1	63.2	39.1	18.0
1909	58.7	147.5	67.0	74.7	54.9	62.5	40.2	17.8
1910	60.2	145.6	68.7	76.4	55.2	63.0	38.7	17.8
1911	59.7	144.9	68.3	79.6	55.9	63.9	38.9	18.6
1912	62.3	132.4	71.4	78.2	57.5	65.9	34.1	17.5
1913	62.6	135.3	71.9	79.7	57.9	66.5	34.6	17.7
1914	63.5	149.7	71.8	78.6	58.2	65.8	37.8	17.5

of the total gold stock.²² Assume further that the world price level rose in proportion to the increased stock of gold. We then have

$$\text{RPGH} = \text{RPG} \cdot \frac{\text{EWMG} + \text{WNMG}}{\text{WMG} + \text{WNMG}}. \quad (\text{A2})$$

Since the real price of silver is by assumption one-sixteenth of the real price of gold and by definition equal to the nominal price (=legal price) divided by the price level, we have

$$\text{PH16} = 1.2929 \dots \cdot \frac{16}{\text{RPGH}}. \quad (\text{A3})$$

The actual U.S. monetary gold stock became a steadily increasing fraction of the world's monetary gold from 1879 on, so that the hypothetical price roughly parallels the actual price, with the differential rising somewhat over the period (see fig. 3 of the text). In 1876, when resumption on silver would have taken place, the price level as estimated from equation (A3) was a trifle below the actual price level. By 1877, it was a trifle above.

The hypothetical real price of gold is also all that is needed to estimate the effect on the price level of the gold standard world if the United States were on a silver standard throughout the period. If we take the U.K. price level as representative of the price level of the gold standard world, we have

$$\text{UKPH} = \text{UKP} \cdot \frac{\text{WMG} + \text{WMNG}}{\text{EWMG} + \text{WNMG}} \quad (\text{A4})$$

(see fig. 4 of the text). The effect is clearly appreciable.²³

C. A More Sophisticated Estimate

To go beyond these simple estimates requires finding a way to estimate the real price of silver since we can use the counterpart of equation (A3) to convert such an estimate into an estimate of the hypothetical price level.

The real price of silver is determined by (a) the supply of and (b) the demand

²² I owe this approach to Hugh Rockoff. It replaces a less attractive assumption I had made initially.

²³ An interesting check on these estimates, discovered after they were completed, is provided by Irving Fisher, who, in 1911, wrote the following: "If some way had been contrived by which gold and silver could have been kept together (say by world-wide bimetallism), prices would not have fallen so much [from the average of 1873-76] in gold countries, or risen so much (if at all) in silver countries, but would probably have fallen in gold countries slightly—probably about 10 per cent up to 1890-1893 and more up to 1896" (pp. 244-45). He estimates that prices in fact fell 22 percent in gold countries between 1873-76 and 1890-93 and rose 17 percent in silver countries. According to table A1, prices in the United States, a gold country, did fall 22 percent between the indicated dates but in the United Kingdom, which I have taken as a representative gold country, by 14 percent. The estimate of the hypothetical price index for the United Kingdom falls by half as much or 7 percent and then falls further to 1896, in both respects very close to Fisher's estimates, especially with respect to the fraction of the decline that would have been avoided. As to silver countries, the estimate of the 16 to 1 U.S. price level in table A1 falls 4 percent and that of the sophisticated price level (discussed below) rises 4 percent, consistent with Fisher's "if at all."

for silver for nonmonetary use in the world as a whole. The adoption of a bimetallic or silver standard by the United States would presumably not have significantly affected the world demand function for silver for nonmonetary use. To estimate that demand function (subsection 3 below) requires data on the actual nonmonetary use of silver (subsection 1). On the other hand, the adoption of a bimetallic or silver standard by the United States would clearly have altered significantly the supply of silver for nonmonetary use (subsection 2) because it would have increased the monetary demand for silver. Constructing acceptable estimates for the period in question (1875–1914) proved by far my most troublesome problem.

1. Actual Nonmonetary Use of Silver

The supply of silver for nonmonetary use is equal to (1) the production of silver minus (2) the demand for silver for monetary use by the rest of the world minus (3) the demand for silver for U.S. monetary use, or

$$\text{SNM} = \text{SPROD} - \text{EWMDS} - \text{UMDS}. \quad (\text{A5})$$

Estimates for SPROD, the annual production of silver, and EWMDS, the increment in the monetary stock of silver by other countries, are readily available. I have constructed estimates for UMDS, the increment in the U.S. monetary stock of silver, for fiscal years 1873–94 from a report of the Treasury Department reporting purchases by the United States under the successive silver purchase acts, and for later years from estimates of the total dollar value of the monetary stock of silver.

2. Hypothetical Supply of Silver for Nonmonetary Use

Equation (A5) gives actual nonmonetary use. Add an H to the relevant symbols and it gives hypothetical nonmonetary use under a silver standard. Item 1, silver production, depends in principle on the real price of silver. However, during the period in question, the actual production of silver rose sharply, nearly tripling from 1880 to 1914, while at the same time the real price of silver fell to less than half its initial level. Supply was clearly being driven by exogenous discoveries and innovations. Moreover, much silver is a by-product of the mining of other metals and so is relatively inelastic in supply. Hence, I have assumed that silver production would have been what it actually was. This assumption introduces an error leading to an upward bias in the estimated real price of silver.

With regard to item 2, I have assumed that other countries would not have been affected by the U.S. adoption of a silver standard, by their either adopting silver rather than gold or changing the amount of silver added to their monetary stocks. This assumption seems eminently justified. The major move from silver to gold by Germany, France, and others came before the United States would have moved to a silver standard, and indeed was part of the reason why the United States itself moved to a gold standard. Hence, I have simply used the actual monetary demand by other countries as the hypothetical estimate.

Item 3, the hypothetical increment in the U.S. monetary stock of silver, is the most difficult. We can tautologically express the hypothetical U.S. monetary silver stock (in ounces) as the product of the ratio maintained between specie and money (SPR) times the quantity of money divided by the legal price of

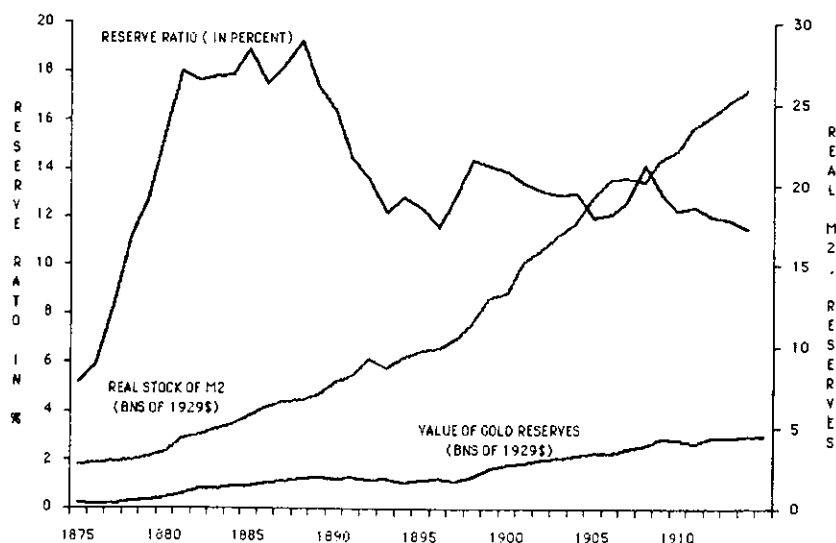


FIG. A1.—Gold reserve ratio, real money stock, and gold reserves in billions of dollars, 1875–1914.

silver or, expressing the quantity of money by the ratio of nominal income to velocity and nominal income as the product of real income and the price level, as follows:

$$\text{UMSH} = \frac{\text{UMG\$}}{\text{UM}} \cdot \frac{y}{V} \cdot \frac{P}{LP} = \text{SPR} \cdot \frac{y}{V} \cdot \frac{1}{\text{RPSH}} = k_1 \cdot \frac{1}{\text{RPSH}}, \quad (\text{A6})$$

where y/V is the real money stock; multiplication by P converts it into nominal dollars. Only the product of SPR and y/V , which I have designated by k_1 and which equals the real value of the specie reserve, enters in the subsequent analysis. (In principle, all the symbols should be followed by an H , but since no confusion arises except for the real price of silver, I have omitted it.)

The reason for expressing the money stock as the product of the real stock and the price level is that the price level is what we are seeking to estimate. The second form of stating the right-hand side of equation (A6) introduces the hypothetical real price of silver in place of the nominal price level. From that, we can readily estimate the hypothetical nominal price level by using the counterpart of equation (A1).

In computing the actual values in equation (A5), I regarded silver in circulation or held by the Treasury as “monetary silver.” However, in estimating hypothetical values of the specie reserve ratio and of specie reserves for the gold standard period, I cannot treat monetary silver as part of specie reserves, though it would have had that status under a bimetallic or silver standard. It was simply a governmental asset accumulated as part of an attempt to prop up the price of silver (like government stocks of wheat at present).

Accordingly, I have used only monetary gold stocks for the present purpose. Figure A1 plots the gold reserve ratio (the ratio of the dollar value of monetary gold to the quantity of money), the real value of the stock of money, and the

real value of gold reserves (actual gold k_1). The rapid rise in the reserve ratio during the first 5 years after the passage of the Resumption Act (1875–79) was to be expected in preparation for resumption. Presumably, a similar rise would have occurred if resumption had been on silver instead of gold, with the sole difference that silver would have been accumulated rather than gold. In either case, the accumulation of reserves required a surplus on the current account of the balance of payment or capital inflows. A sizable surplus was generated from 1876 to 1881 followed by sizable capital inflows. I see no reason to suppose that the initial buildup of reserves would have been different under silver from what it was under gold.

By 1879, the specie reserve ratio reached roughly the same level as in the early 1900s after the end of the period of uncertainty generated by the monetary disturbances of the 1880s and 1890s. The further rise after 1879 was prompted by an effort to persuade the public, not only at home but equally abroad, that the gold standard was here to stay. As agitation for a more expansive monetary policy mounted, however, that effort failed and, especially after the prosilver movement gained steam, led to continuous pressures on gold reserves, producing a sharp decline in the reserve ratio and a slightly declining level of real reserves. After the defeat of Bryan in 1896, there was a temporary spurt in the reserve ratio and an even sharper rise in real reserves as the higher reserve ratio was reinforced by a rapid increase in the real money supply, itself partly a consequence of a return of confidence that both lowered velocity and fostered a higher real income. Reasonably steady conditions were not attained until the end of our period.

After trying many alternative ways of estimating what specie reserves would have been under an unchallenged and fully accepted silver standard, I finally settled on a purely empirical expedient: a straight-line trend between the average values of gold reserves during the first 5 and the last 5 years of the period from 1875 to 1914. As figure A2 shows, such a trend eliminates both the initial bulge and later decline that I attributed in the prior paragraph to the monetary disturbances and their aftermath. For 1875–79 and 1901–14, it approximates the actual pattern.

The U.S. hypothetical annual monetary demand for silver is simply the increment in the U.S. hypothetical silver stock:

$$\text{UMDSH}(t) = \Delta \text{UMSH} = \text{UMSH}(t) - \text{UMSH}(t-1). \quad (\text{A7})$$

The possible errors in this approach are numerous. Some simply affect the year-to-year movements as a result of the use of a trend for k_1 . Any systematic bias arises primarily from the assumption that the same specie reserves would have been maintained under a silver standard in the early and late years of the period as those maintained under a gold standard. The possible sources of error are different for the specie reserve ratio and the real stock of money. The desired specie reserve ratio might have been affected by a different pattern of prices. Rising prices under a gold (silver) standard means that the real value of gold (silver) is falling, and conversely. A falling real value makes it less expensive to hold the specie reserves, and conversely. It is doubtful, however, that any such price effect has a significant influence on the decision by the monetary authority on how large a specie reserve is desirable: any financial benefit or loss is subtle and accrues to the government at large and not specifically to the monetary authority. A more important factor is surely the threat of a specie drain, which would have been largely absent under a secure silver standard.

The real stock of money would have been affected by the reduction in un-

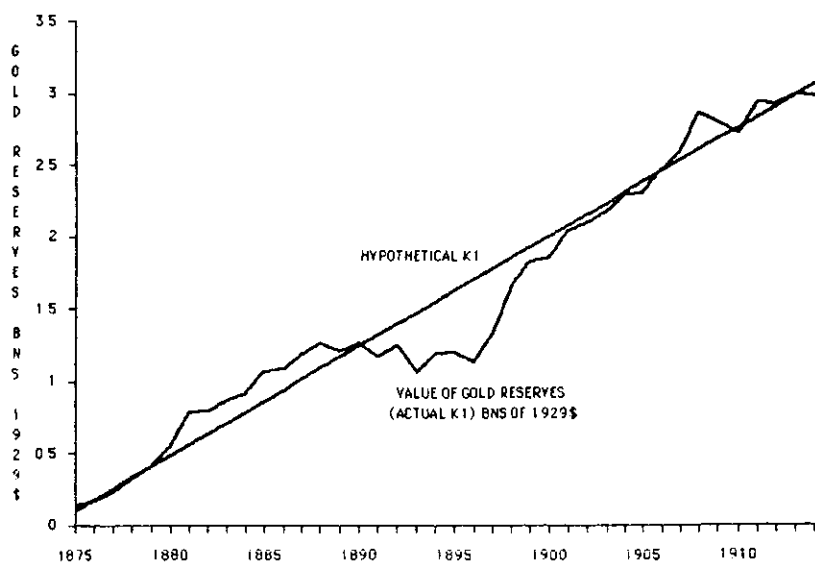


FIG. A2.—Actual and hypothetical gold reserves (k_1) in billions of 1929 dollars, 1875–1914.

certainty as a result of settling on a definitive silver standard. The reduced uncertainty would have tended to lower velocity and raise real income, both of which would have raised the real money stock—as appears to have happened after 1896. Neglect of these effects produces an underestimate of the hypothetical silver stock. Such an underestimate introduces a downward bias in estimating the real price of silver, or a bias in the opposite direction from the possible bias introduced in estimating item 1, the production of silver.

3. Demand for Silver

The quantity of silver demanded for nonmonetary use depends primarily on world real income, the real price of silver, and the real price of gold. I have estimated a demand curve with these variables in two variants: linear and logarithmic. In general, the logarithmic form is preferable. However, in this particular case I do not believe it is. The logarithmic form forces the nonmonetary demand for silver to be positive, yet it is easily possible for additions to monetary stocks of silver to exceed the world production of silver, as happened most recently under the silver purchase program of President Roosevelt in the 1930s. In that case, the quantity of silver available for nonmonetary purposes is negative if it is estimated in accordance with equation (A5), which gives the nonmonetary supply of silver out of current production, not the nonmonetary use of silver.

As an estimate of world real income, I have used an index number of the

physical volume of world production given by Warren and Pearson (1933).²⁴ For the real price of silver and the real price of gold, I simply used the actual prices divided by the U.S. deflator. This procedure assumes that the real price of silver and the real price of gold were the same throughout the world, surely a not unreasonable assumption for those two monetary metals.²⁵

The two equations are as follows for 1880–1914:

$$\log \text{SNM} = \underset{(3.7)}{-6.96} + \underset{(4.0)}{1.27} \log \text{WI} - \underset{(4.0)}{1.28} \log \text{RPS} + \underset{(5.6)}{1.87} \log \text{RPG}, \quad (\text{A8})$$

$$\text{SNM} = \underset{(.8)}{58.28} + \underset{(4.0)}{2.13} \text{WI} - \underset{(4.0)}{66.21} \text{RPS} + \underset{(1.1)}{.88} \text{RPG}. \quad (\text{A9})$$

where WI stands for world income. As usual, the values in parentheses are the absolute *t*-values. In the log equation, the coefficients are all highly significant; in the linear equation, only the coefficients of world income and the real price of silver are. However, there is little to choose between the equations in terms of the goodness of fit, as can be seen graphically in figure A3 as well as from the adjusted *R*²s which are .949 for the log equation and .950 for the linear equation. The standard error of estimate for the log equation is .180, which is comparable to an estimate of the coefficient of variation for the linear equation. That is .138 if the denominator of the coefficient of variation is the arithmetic mean of the dependent variables, and .177 if it is the geometric mean. Both estimates for the linear equation are lower than that for the log equation.

Estimating a hypothetical price level using the linear equation is mathematically far more tractable than using the log equation, which reinforces the theoretical consideration in favor of the linear equation, that is, that the silver available for nonmonetary use out of current production can be negative. Hence, from here on, I use only the linear equation.

4. Equating Supply and Demand

Equating equations (A5) and (A9) and rearranging terms yield

$$\begin{aligned} \text{UMDSH} = & \text{SPROD} - \text{EWMDS} - 58.28 - 2.13\text{WI} \\ & - .88\text{RPGH} + 66.21\text{RPSH}. \end{aligned} \quad (\text{A10})$$

²⁴ A footnote to the Warren and Pearson tables says that for 1865–1932 the index was prepared by Carl Snyder of the Federal Reserve Bank of New York. Warren and Pearson report similar index numbers of U.S. physical volume of production. The trend of the index of U.S. production is steeper than the trend of U.S. real income (estimates from Friedman and Schwartz [1982]). On the other hand, the general ups and downs are very similar. Accordingly, I experimented with adjusting the Warren and Pearson index by subtracting out a trend at a rate equal to the difference between the logarithmic trends of U.S. production and U.S. real income, which was 0.4 percent a year. However, the effect on the final results was trivial and, if anything, rendered them slightly less significant statistically, so I have simply used the original index.

²⁵ However, it is not clear that it is preferable to use the U.S. rather than the U.K. deflator. I experimented with both. The difference in results was small, trivially in favor of the U.S. deflator. However, a more decisive consideration is that I want to use the equation to estimate the hypothetical U.S. price level, so it was encouraging that substituting the U.K. deflator did not produce a statistical improvement.

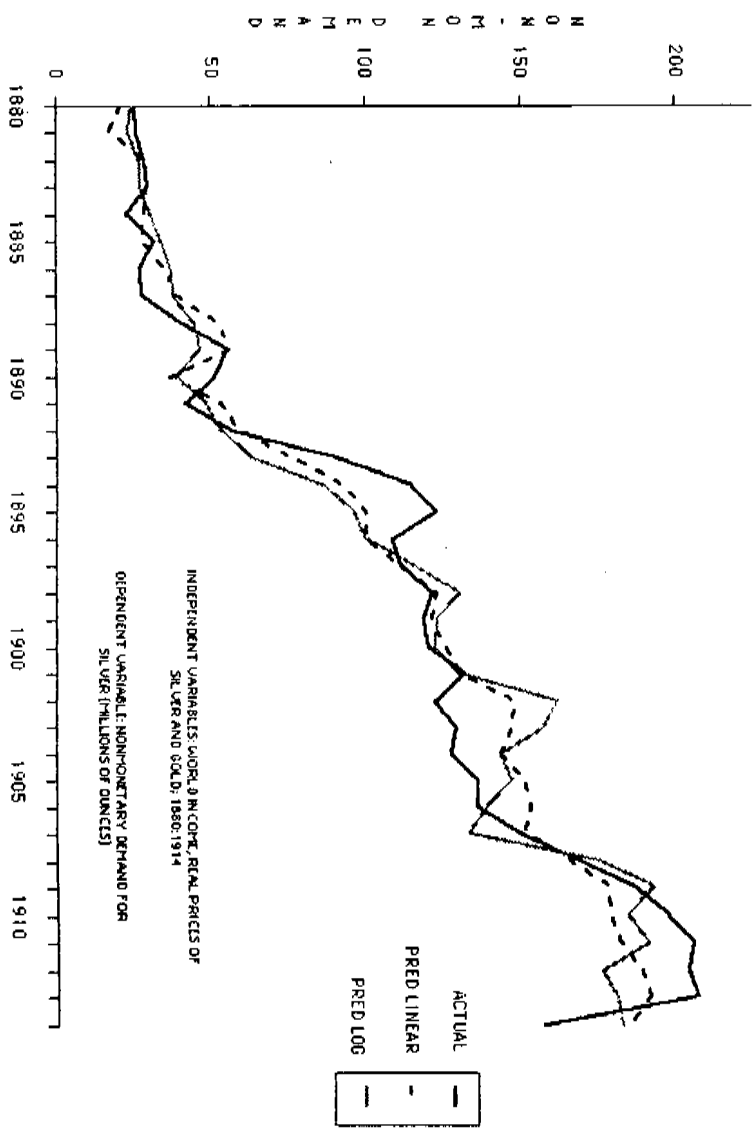


Fig. A3.—Nonmonetary demand for silver: actual vs. predicted by linear and log regressions.

To simplify, let k_2 equal all the terms on the right-hand side of (A10) except the last, and let x equal the hypothetical real price of silver that is our objective. All these are also functions of time. However, given assumptions up to this point, we have estimates of the values of k_1 and k_2 for all the years from 1874 to 1914.

In terms of these symbols, we can rewrite equation (A7), using equation (A6), as

$$\text{UMDSH}(t) = \frac{k_1(t)}{x(t)} - \frac{k_1(t-1)}{x(t-1)}. \quad (\text{A11})$$

Equating equations (A10) and (A11) and simplifying yield

$$66.21x^2(t) + \left[k_2(t) + \frac{k_1(t-1)}{x(t-1)} \right] x(t) - k_1(t) = 0. \quad (\text{A12})$$

Equation (A12) is now in the form of a straightforward quadratic equation except for the troublesome presence of the term including $x(t-1)$ in the denominator; $x(t-1)$ is one of the unknowns that we are trying to determine. As a first approximation, assume that the real price of silver does not change from year to year, that is, that $x(t)$ equals $x(t-1)$. That assumption converts equation (A12) into the simplified equation (A13), which involves only the current year's value of the unknown x , although it does involve the prior year's value of k_1 , via substituting Δk_1 for $k_1(t) - k_1(t-1)$:

$$66.21x^2 + k_2x - \Delta k_1 = 0. \quad (\text{A13})$$

The solution to this equation is a first approximation to x .

For a second, third, and so forth approximation, we can return to equation (A12) and replace $x(t-1)$ with the prior approximation estimate. The successive approximations do converge, though rather slowly. The main changes are not in the level or general pattern but rather in the year-to-year movements. However, each approximation involves losing one value at the beginning of the series. I stopped with the eleventh approximation, at which point 1884 is the first year for which there is an estimate. For earlier years, I used the earlier approximations, beginning with the third for 1876, the year in which the silver standard would have been adopted.²⁶ Given this estimate of the real price of silver, it is necessary only to divide the legal price by the real price to estimate the hypothetical price level under a silver standard. The resulting "sophisticated" estimate of the hypothetical price level for the United States is plotted in figure 3 of the text.

5. Gold-Silver Price Ratio

Since we have already estimated the hypothetical price of gold, it is trivial to get the hypothetical price ratio of gold to silver. The result is plotted in figure 2 of the text along with the actual and legal gold-silver price ratio. The actual gold-silver price ratio under a U.S. silver standard would almost surely have fluctuated much less than our estimates of the hypothetical gold-silver price ratio, given the arbitrary assumptions and inevitable measurement errors that affect our estimates and the extent to which they have been affected by the monetary uncertainty of the period.

²⁶ The first year for which I have an estimate for EWMS is 1874, which explains why the first year for which I can estimate the first approximation is 1874.

These estimates suggest that, if the United States had returned to a bimetallic standard in 1879 and had stayed on it consistently throughout, the market gold-silver price ratio would have remained roughly equal to or only slightly above the U.S. legal price ratio, just as for close to a century it remained roughly equal to the legal price ratio in France. (Table A1 gives the numerical values for the curves plotted in figs. 2, 3, and 4 of the text.)

D. An Even More Sophisticated Estimate

In principle, it would be possible to get a fully simultaneous solution for both the real price of silver and the real price of gold by following the same procedure for gold as for silver, namely, estimating (a) a demand equation for the non-monetary use of gold and (b) the hypothetical quantity of gold that would have been demanded if the United States had gone on a silver standard. However, explorations along this line proved disappointing. First, estimating the hypothetical nonmonetary demand is even more difficult for gold than for silver.²⁷ Second, estimated demand equations for gold yielded negative, though statistically not significant, coefficients for the real price of silver instead of the positive coefficient that would be expected for a substitute for gold.²⁸ That result is inconsistent with the positive coefficient for the real price of gold in demand

²⁷ A major sticking point is specifying precisely how the U.S. gold stock would have been disposed of. My earlier rough approximation evades this question. For a full solution, however, we cannot do so. Demand functions for gold and silver refer to annual quantity demanded, and we need to equate that demand function with annual supply. This means that we would need to add to total gold production the amount of gold that the United States would have released to the rest of the world from its stock on a year-by-year basis. I see no way to estimate the annual release except by purely arbitrary assumptions.

²⁸ The calculated demand functions for nonmonetary use of gold are as satisfactory as those for silver in terms of goodness of fit, but not in terms of economic logic. The logarithmic and linear demand functions are as follows:

$$\log \text{WGNM} = \frac{4.34}{(13.2)} + \frac{.555 \log \text{WI}}{(10.0)} - \frac{.077 \log \text{RPS}}{(1.4)} - \frac{.259 \log \text{RPG}}{(4.5)}, \quad (\text{A14})$$

$$\text{WGNM} = \frac{169.862}{(4.2)} + \frac{3.08 \text{WI}}{(9.0)} - \frac{8.721 \text{RPS}}{(.8)} - \frac{1.482 \text{RPG}}{(2.8)}, \quad (\text{A15})$$

where WGNM is the world nonmonetary demand for gold. As with silver, both equations give high multiple correlations (adjusted R^2 's of .98 for the log equation and of .97 for the linear equation) and relatively small standard errors. The standard error of the log correlation is .031. The corresponding estimate for the linear equation of the coefficient of variation is .037 whether the denominator is an arithmetic mean or a geometric mean. An appendix to chap. 4 of the report of the U.S. Commission on the Role of Gold (1982, pp. 176–77) reports estimates of demand equations linear in the logarithms of the variables for the industrial demand for gold for 1950–80 and 1969–80. The independent variables are conceptually the same as those that I used: the real price of gold, the real price of silver, and real income. Both sets use two alternative deflators to estimate the real prices: the U.S. wholesale price index and the world consumer price index. The difference between the two sets of equations is that the one for the longer period uses U.S. income only, whereas that for the shorter period uses three alternative real income variables: for seven major industrial countries, for the United States, and for the world. The four equations that use U.S. income give a negative coefficient for the real price of silver, though only one out of four comes close to statistical significance. On the other hand, the four others (all for the shorter period) are all positive, in line with theoretical expectations, though none comes close to statistical significance. This evidence clearly does not contribute to resolving the puzzle.

equations for silver, a violation of the Slutsky cross-equation condition. Eliminating that contradiction requires estimating the silver and gold demand equations simultaneously, imposing the appropriate cross-equation restriction. Given the drastic difference between the two separate equations, I doubt that the result would deserve much confidence. Finally, these problems resolved, the simultaneous solution requires solving a fourth-degree equation in the U.S. price level.

I am thus left in a quandary: I am unhappy with what I have done, but even more unhappy with the most obvious alternative, a simplified general equilibrium analysis. A comprehensive general equilibrium analysis would have to include not only the determinants of gold production and silver production, which I have completely neglected, but also the determinants of the fraction of gold production and silver production that go into monetary and nonmonetary use. Construction of such an expanded general equilibrium model would be extremely laborious and would deserve little confidence. Under the circumstances, I am inclined to leave well enough alone, while at the same time acknowledging that the estimates are subject to a wide margin of error, particularly with respect to year-to-year movements.

Appendix B

A. *Source of Variables*²⁹

<i>P</i>	1869–1914: Friedman and Schwartz (1982, table 4.8); 1865–68: extrapolated backward from Hoover (1960, p. 142).
PS	U.S. Bureau of the Census (1975, p. 606, ser. 270). For 1865–78, original source gives price in gold dollars. Adjusted to greenback price by multiplying by reciprocal of gold value of currency, from Warren and Pearson (1933, p. 351, table 69).
WMG, WNMG	U.S. Commission on the Role of Gold (1982, p. 198, table SC 7).
UMG	U.S. Commission on the Role of Gold (1982, p. 203, table SC 9); 1865–77 shifted from June 30 to December 31 data by 2-year moving average of June 30 data for 1865–78.
EWMG	Equals WMG – UMG.
RPG	1865–78: reciprocal of gold value of currency (from Warren and Pearson [1933, p. 351, table 69]) times legal price of gold (\$20.67183) divided by <i>P</i> ; 1879–1914: legal price divided by <i>P</i> .
UKP	1868–1914: Friedman and Schwartz (1982, table 4.9); 1865–67: extrapolated back from 1868 by price index implicit in Deane (1968).
SPROD	Warren and Pearson (1933, p. 139, table 24). 1865–75: linear interpolation between centered 5-year averages in table. Thereafter, annual figures in table.
EWMDs	Drake (1985, pp. 208–9, table A) gives estimates for successive 5-year periods based on annual reports of the director of the mint. I simply assumed that the same amount was accumulated each year during the successive 5-year periods. The numbers are small and do not vary drastically from one period to the next, so not much error is introduced by this assumption. How-

²⁹ Sources are listed in the order in which the variables are introduced in App. A.

ever, I suspect that the initial estimates are subject to a large margin of error.

UMDS	Purchases under the silver purchase laws of February 12, 1873; January 14, 1875; February 28, 1878; and July 14, 1890, are given in U.S. Secretary of the Treasury (1899, p. 207). For the first two purchase laws only the total is given; I have assumed that the amount purchased was the same in each month of the period for which each law was in effect. For the final two laws, figures are given for fiscal years from 1878 to 1894. For later years, I have estimated the physical stock of silver from the dollar stock of silver dollars and subsidiary silver as reported in U.S. Secretary of the Treasury (1928, pp. 552–53) by dividing by the legal price, allowing roughly for the lesser amount of silver in subsidiary silver, and differenced the series to get annual purchases. When the Treasury bought silver, it paid the market price but valued it for monetary purposes at the legal price of silver, which is why the physical stock can be estimated from the monetary stock by dividing by the legal price of silver. The allowance for the different treatment of subsidiary silver is rough, but the amounts involved are small, so no great error is introduced. The final estimates are for fiscal years ending June 30, whereas SPROD and EWMS are for calendar years, so I have converted the fiscal year data to calendar year data by a 2-year moving average.
UMG\$	1879–1914: Friedman and Schwartz (1963, pp. 130–31, table 5); 1866–78: estimates by Anna J. Schwartz based on same sources.
UM	Friedman and Schwartz (1982, table 4.8).
y	Friedman and Schwartz (1982, table 4.8).
V	Equals nominal income from Friedman and Schwartz (1982, table 4.8) divided by UM.
WI	Equals Warren and Pearson (1933, pp. 85–86, table 12) index number of world's physical volume of production, 1880–1914 = 100, divided by two.

B. Record of Notation

EWMS	Actual monetary demand for silver in the rest of the world (external)
EWMS	Rest-of-world actual monetary gold stock
k_1	Equals $SPR \cdot (y/V)$
k_2	Equals $SPROD - EWMS - 58.28 - 2.13WI - .88RPGH$
LP	Legal price of silver
P	U.S. price level
PHN	Naive estimate of hypothetical price level
PH16	Hypothetical price level on assumption that the gold-silver price ratio is 16 to 1
PS	Nominal price of silver
RPG	Real price of gold
RPGH	Hypothetical real price of gold
RPS	Real price of silver in 1929 dollars
RPSH	Hypothetical real price of silver in 1929 dollars

RPSH16	Hypothetical real price of silver on 16 to 1 ratio assumption
SNM	Silver available for nonmonetary use
SPR	Specie reserve ratio
SPROD	Total silver production
UKP	British price level
UKPH	Hypothetical British price level
UMDS	Actual annual monetary demand for silver in the United States
UMDSH	Hypothetical U.S. annual monetary demand for silver
UMG	U.S. monetary gold stock in ounces
UMG\$	U.S. monetary gold stock in dollars
UMGR\$	U.S. monetary gold stock in 1929 dollars
UMS	Actual U.S. monetary stock of silver
UMSH	Hypothetical U.S. monetary stock of silver
V	U.S. velocity
WI	World real income (including the United States)
WMG	World monetary gold
WNMG	World nonmonetary demand for gold (including the United States)
x	Equals RPSH
y	U.S. real income

References

- Barnett, Paul S. "The Crime of 1873 Re-examined." *Agricultural Hist.* 38 (July 1964): 178–81.
- Bryan, William Jennings. *The First Battle: A Story of the Campaign of 1896*. Chicago: Conkey, 1896.
- Deane, Phyllis. "New Estimates of Gross National Product for the United Kingdom, 1830–1914." *Rev. Income and Wealth* 14 (June 1968): 95–112.
- Drake, Louis S. "Reconstruction of a Bimetallic Price Level." *Explorations Econ. Hist.* 22 (April 1985): 194–219.
- Fisher, Irving. *The Purchasing Power of Money: Its Determination and Relation to Credit, Interest and Crises*. New York: Macmillan, 1911.
- Friedman, Milton. "Bimetallism Revisited." *J. Econ. Perspectives* 4 (Fall 1990).
- Friedman, Milton, and Schwartz, Anna J. *A Monetary History of the United States, 1867–1960*. Princeton, N.J.: Princeton Univ. Press (for NBER), 1963.
- . *Monetary Trends in the United States and the United Kingdom: Their Relation to Income, Prices, and Interest Rates, 1867–1975*. Chicago: Univ. Chicago Press (for NBER), 1982.
- Hoover, Ethel D. "Retail Prices after 1850." In *Trends in the American Economy in the Nineteenth Century*. Studies in Income and Wealth, vol. 24. Princeton, N.J.: Princeton Univ. Press (for NBER), 1960.
- Jastram, Roy W. *Silver: The Restless Metal*. New York: Wiley, 1981.
- Laughlin, James Laurence. *The History of Bimetallism in the United States*. New York: Appleton, 1886; 2d ed., 1896.
- Linderman, Henry R. *Money and Legal Tender in the United States*. New York: Putnam, 1877.
- National Executive Silver Committee. *Silver in the Fifty-first Congress*. Washington: Gray, 1890.
- Nugent, Walter T. K. *Money and American Society, 1865–1880*. New York: Free Press, 1968.

- O'Leary, Paul M. "The Coinage Legislation of 1834." *J.P.E.* 45 (February 1937): 80–94.
- . "The Scene of the Crime of 1873 Revisited: A Note." *J.P.E.* 68 (August 1960): 388–92.
- Reagan, John H. Testimony. *Congressional Record*. 51st Cong., 1st sess., 1890. Vol. 21, pt. 3.
- Ricardo, David. *Proposals for an Economical and Secure Currency*. 1816. In *The Works and Correspondence of David Ricardo*, vol. 4, *Pamphlets and Papers, 1815–1823*, edited by Piero Sraffa. Cambridge: Cambridge Univ. Press (for Royal Econ. Soc.), 1951.
- . "Evidence on the Resumption of Cash Payments." Testimony before a committee of Parliament. 1819. In *The Works and Correspondence of David Ricardo*, vol. 5, *Speeches and Evidence*, edited by Piero Sraffa. Cambridge: Cambridge Univ. Press (for Royal Econ. Soc.), 1952.
- Rockoff, Hugh. "The 'Wizard of Oz' as a Monetary Allegory." *J.P.E.* 98 (August 1990): 739–60.
- Rolnick, Arthur J., and Weber, Warren E. "Gresham's Law or Gresham's Fallacy?" *J.P.E.* 94 (February 1986): 185–99.
- Stewart, William M. "Silver the Money of the People." In *Papers and Addresses before the First National Silver Convention*, edited and compiled by E. A. Elliott. St. Louis: Buxton & Skinner, 1889.
- U.S. Bureau of the Census. *Historical Statistics of the United States, Colonial Times to 1970*. Bicentennial ed. Washington: Government Printing Office, 1975.
- U.S. Commission on the Role of Gold in the Domestic and International Monetary Systems. *Report to the Congress*. Vol. 1. Washington: Government Printing Office, March 1982.
- U.S. Secretary of the Treasury. *Annual Report of the State of the Finances for the Fiscal Year Ended June 30, 1899*. Washington: Government Printing Office, 1899.
- . *Annual Report for 1928*. Washington: Government Printing Office, 1928.
- Walker, Francis A. "The Free Coinage of Silver." *J.P.E.* 1 (March 1893): 163–78.
- . *International Bimetallism*. New York: Holt, 1896.
- Warren, George F., and Pearson, Frank A. *Prices*. New York: Wiley, 1933.

Afterword

Milton Friedman as a Microeconomist

I will concentrate on Milton Friedman as a microeconomist. He made major contributions to this field—as I will try to indicate—and his most famous course at Chicago was by far the two-quarter graduate sequence on price theory. Bob Lucas, I, Sherwin Rosen, Gene Fama, and legions of others who went through Chicago to study economics while he taught there remember that course as having a major impact on our approach to economics.

I. Foundation of His Approach

Friedman's approach to microeconomics differed greatly from what was typical then, or what is typical even today. Before giving several illustrations, I will delineate several assumptions he made in his research and teaching on micro.

1. Friedman did not draw a sharp distinction between micro and macro in terms of the tools of economics to be used. He went rather seamlessly from one to the other, as in his discussion of the returns on education, the consumption function, or the natural rate of unemployment.

2. Closely related, Friedman did not conceptually draw a distinction between partial and general equilibrium analysis. So-called partial equilibrium analysis of, say, the demand for cars took account of effects in other markets, such as the gasoline or bus markets, to the extent that changes in these markets were relevant to what occurred in the car market. This is what I call "empirically driven" general equilibrium analysis. Similarly, the usual so-called general equilibrium analysis ignores many potential effects in various markets that are assumed not to have major return effects on the markets one is analyzing. These ignored effects usually include induced changes in tax rates and government

Based on remarks presented at the American Economic Association meetings, January 5, 2007.

spending levels, endogenous changes in preferences, and endogenous changes in technology.

3. I have used the term “market” for good reason since Friedman’s approach to micro as well as macro emphasized markets and market equilibrium rather than the individual and individual choices. To be sure, he started usually with individual choices because he was confident that this gave insights into how markets respond to various stimuli. He also recognized that not everyone maximized perfectly, and some even made *ex ante* mistakes. But he believed that these would greatly cancel out in understanding how markets performed. On the whole, I believe that is the right position in the great majority of occasions.

4. Friedman was not interested in theory for its own sake, but as a means of explaining behavior in implicit and explicit markets. So his theory was oriented toward explaining what happened in the real world. After all, he was a superb statistician for his time and an excellent analyzer of data, and he built on these skills in his analytical work on economic theory.

5. Friedman was convinced that in order to prescribe what government policy should be, one first had to understand how markets actually behaved, including how participants responded to changes in their incentives. But once he felt he did understand these responses, he was not shy about suggesting improvements in what governments should do!

II. Some “Micro” Contributions

All these characteristics are revealed by Friedman’s work in different parts of economics. I will try to show this with a brief discussion of some of his important contributions to microeconomics. I start with his first major publication in economics: *Income from Independent Professional Practice* (National Bureau of Economic Research, 1945). Although the book was published jointly with Simon Kuznets, the preface delineates Friedman’s contributions—necessary because this was his Columbia dissertation—and I concentrate on these. This is a pioneering study that illustrates at an early stage of his career virtually all the principles I claim to have guided his research. This book is an absolutely superb empirical study that introduced the concept of permanent income to analyze the stability of a professional’s income over time, was an early human capital analysis of returns to education, and was perhaps the first quantitative study of union restrictions on entry into occupations and of the effect of unions on earnings.

Friedman calculated the returns to entering medicine relative to dentistry and engineering. He concluded that there is a sizable earnings premium to medicine that he attributes to the power of the American

Medical Association in restricting entry into medicine through controls over certification of doctors. Publication of this study by the NBER was held up for several years, and hence so too was Friedman's diploma from Columbia, because a board member of the bureau objected to this conclusion and analysis of the AMA's behavior. Friedman refused to budge, and eventually a compromise was reached whereby the published book included a largely irrelevant comment by the objecting board member.

The concept of permanent income was introduced as a way of estimating the persistent component in individual earnings in the three professions analyzed and how the degree of persistence differed across these professions. He used the covariance structure in individual earnings from each profession over these few years to estimate what individual earnings would be expected to look like in the years beyond those that were available.

Friedman's most important use of the concept of permanent income is of course not in this book, but in his classic *Theory of the Consumption Function* (Princeton University Press, 1957), published 12 years after the professional income book. To me, the consumption function book is his most innovative and complete analysis, which combines forward-looking income behavior that builds on his concept of permanent as opposed to current income with an Irving Fisher type of analysis of the relation between consumption and permanent income and with an econometric analysis of how to match up theoretical concepts with empirical data.

The econometrics is based on an errors-in-variable analysis that builds on Friedman's skills as a statistician. Most economists do not realize that his first important article was not in economics, but studied a purely nonparametric statistical issue: the use of ranks in analyzing data, published in 1937 in the *Journal of the American Statistical Association* ("The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance" [32 (September 1937): 675–701]).

Friedman's analysis of the relation between consumption and income in his book on the consumption function essentially overthrew virtually all the alleged findings to that point on the relation between consumption and income when comparing, say, different occupations, blacks and whites, rural and urban dwellers, and aggregate income and aggregate consumption—the latter based typically on a crude Keynesian-type approach. I was a little involved in some of the backlash to his approach since Friedman and I published an article ("A Statistical Illusion in Judging Keynesian Models" [1957], in this volume) that built on the consumption–permanent income relation to attack the usual Keynesian analysis of this relation and its implications for the Keynesian picture of the world.

I want to say a few words about Friedman's contributions to a very different field: choice under uncertainty and expected utility theory. Friedman and Jimmie Savage published an influential analysis in the late 1940s ("The Utility Analysis of Choices Involving Risk" [1948], in this volume) that used the then recent revival of expected utility theory by John Von Neumann and Oskar Morgenstern to explain why most people both insure and gamble. They ended up with a utility function that is concave in some regions and convex in others. They briefly interpret the move from concave to convex regions as corresponding to movement from lower to higher social classes.

There is obviously an incentive to try to participate in gambles or lotteries that take one out of a convex region because the marginal utility of income is rising in that region. Not long after the Friedman/Savage article was published, Harry Markowitz, a student of Friedman's, introduced the concept of a reference point. This implies that the utility function shifts upward or downward as a person's usual or permanent income position rises or falls. Markowitz also introduced the concept of loss aversion around the reference point, which muted the incentive to gamble. He had other innovations too that greatly influenced subsequent developments of so-called behavioral models of choices under uncertainty.

Friedman himself took a different approach that has been less influential, but I believe is extremely important. In a later article ("Choice, Chance, and the Personal Distribution of Income," *J.P.E.* [61 (August 1953): 277–90]), he used the implication that people in convex regions of their utility functions would take fair and even some unfair gambles to derive one of the first endogenous theories of the personal distribution of income. In his theory, the resulting equilibrium distribution of income is based somewhat on the initial income distribution but is also sensitive to the incentives to take risks that take one out of certain initial income regions. I believe that this approach has many insights and deserves much more investigation. Follow-ups to Friedman's insights into endogenizing the distribution of income include papers by Sherwin Rosen ("Manufactured Inequality," *Journal of Labor Economics* [14 (April 1997): 189–96]) and by Kevin M. Murphy, Iván Werning, and me ("The Equilibrium Distribution of Income and the Market for Status," *J.P.E.* [113 (April 2005): 282–310]).

III. Public Policy

Finally, I will say a little about Friedman's many ingenious and influential policy proposals, including those in *Capitalism and Freedom* (University of Chicago Press, 1962) and in other collections of his policy writings. These proposals chiefly build on two principles that I believe continue

to be highly relevant in understanding behavior. One is the belief based on much evidence that competition in all walks of life generally improves outcomes to consumers. The second is that individuals in the vast majority of situations know their own interests better than government officials or anyone else know them.

Perhaps Friedman's most famous proposal pertains to a system of education vouchers ("The Role of Government in Education," in *Economics and the Public Interest*, edited by Robert A. Solo [Rutgers University Press, 1955]). According to his article, which shocked many when it was first published in 1955, governments should continue to fund kindergarten through grade 12 education, but they should provide parents with vouchers that they could spend on the public or private schools of their choice. In this way, greater competition would be attracted into the education industry that would develop curricula to cater to parents' and students' interests rather than those of teachers or professional educators. Some parts of the world have introduced such vouchers, and they continue to grow slowly in the United States.

Friedman also supported the volunteer army, so that young men and women could choose the occupations they preferred rather than be forced into military or other service because of government fiat. He was clearly the most influential person on the Gates Commission that, in 1970, issued its report. The report successfully advocated replacing the draft by an all-volunteer armed forces.

Friedman was one of the first advocates of a flat income tax as a way of reducing the amount of resources devoted to tax evasion and avoidance, and of reducing the distortions caused by high marginal taxes on the allocation of time between work and leisure. Perhaps, however, he should have worried more about the effect of flat taxes in stimulating higher tax rates precisely because marginal deadweight costs of taxes would then be lower.

Friedman came out for flexible exchange rates against the prevailing wisdom at that time in the 1950s that flexible rates would be unstable. Eventually, the development of forward markets on currencies by the Chicago Mercantile Exchange in the early 1970s, in which Friedman played an important role, proved empirically that flexible exchange rates would generally be a stable and attractive system.

To mention a few other important but controversial positions, Friedman opposed unions not because he thought that they caused inflation—as others did—but because they led to misallocations of labor across activities. He wanted a system of individual personal retirement accounts to replace pay-as-you-go systems of social security because he believed that these systems reduced work and distorted retirement and other decisions. He supported a negative income tax system, such as the earned income tax credits, over welfare systems because he believed

that negative income taxes had fewer distorting effects on the shift between household and market labor.

IV. Conclusions

It is early to assess accurately Milton Friedman's place in the history of economic science. Surely, however, he was one of the most influential economists of the twentieth century. His influence stemmed partly from his deeply analyzed and energetic espousal of free-market policies at a time when socialism was first on the rise and then completely collapsed as government central management of an economy was shown to be a failure in the Soviet Union, China, India, and elsewhere. But his influence also stemmed from his many innovations in economic science. I have surveyed only a small number of these in a few fields, but these few alone show the power and penetration of Friedman's mind and understanding of economics and his relentless pursuit of the connection between economic theory and empirical evidence.

GARY S. BECKER

Index

Page numbers in *italics* refer to figures or tables.

A

absolute vs. relative prices, 17
aggregate nominal demand, and
 changes in unemployment and
 price levels, 4
alternative consumption functions
 errors in predicting investment,
 82–85
 interpretation of results, 85–87
 MF conclusions on, 87–88
 statistical interpretation, 79–82
 theories, 72–79, 78
Armstrong, W. E., 34n
Ashton, John, 33n
autonomous consumption, 72, 74n

B

Bailey, Martin J., 128n, 129, 135, 140, 143
Barnett, 147n
Baumol, William J., 57, 58–59, 60, 60n, 61, 65
 on expected-utility hypothesis, 65–67, 68, 69
 “The Neumann-Morgenstern Utility Index—an Ordinalist View,” 59n
Becker, Gary, 1n, 9–10, 118n
 on Keynesian models, 72–88
 on MF place in history, 186
Bernoulli, Daniel, 25, 25n, 57, 59n
Bernoulli, Nicholas, 59n
Biddle, Nicholas, 149
bimetallic standard. *See* Coinage Act of 1873
 effect on price of goods and services, 156–158

 outside U.S., 155
bimetallism, 164–172, 168n
Bordo, Michael D., 146n
Braun, Conrad, 146n
Brill, Daniel, on saving and investing propensities, 124, 124n, 129, 129–30n, 132
Brunner, Karl, 1n
Bryan, William Jennings, 146, 149, 163, 171

C

Cagan, Phillip, 1n, 72n, 114n, 128n, 139n, 140, 146n
Callaghan, James, 10
Cannan, Edwin, on investment ROI, 30
Chicago Mercantile Exchange, 185
Christ, Carl, 72n
Cobb, Joe, 146n
Coinage Act of 1873
 consequences of, 154–58
 as “crime,” 152–54
 historical background, 146–52
consumption function, 72–88
Cramer, Gabriel, 25n, 57, 59n

D

demand for money and interest rates, 132–33
Doblin, Ernest, 89n
Drake, Louis, 164n
du Pont, Pierre S., 21

E

economics, as science, 1–3
education
 value of, 181, 182
 vouchers, 185
elasticity and interest, 120, 132

employment and productivity, 8–9,
14
Ewen, C. L'Estrange, 31–32n
exchange rates, 185
expected utility hypothesis
background, 57–59
evidence for acceptance, 59–65
measurement of utility, 65–70
MF conclusions on, 70–71

F

Fand, David, 120n
Federal Reserve control over
available money, 94
fiscal policy, zero interest elasticity
and, 132
Fisher, Irving, 25, 69, 116, 116n, 131,
151n, 168n, 183
Fisher, R. A., 2
flat income tax, 185
flexible exchange rates, 185
Ford, James, 103n
Friedman, David, 146n
Friedman, Milton, opinions and
commentary
on alternative consumption
functions, 87–88
on bimetallism, 164–72
on consumption function, 72–88
on “crime of 1873,” 159–64
on education vouchers, 185
on expected utility hypothesis, 70–
71
on flat income tax, 185
on flexible exchange rates, 185
on inflation, 136–40, 137n
on natural rate hypothesis, 6–9
on people's thoughts about
money, 158
on unions, 185
on volunteer army, 185
Friedman, Milton, works
*A Monetary History of the United
States, 1867–196.*, 121n, 122
A Program for Monetary Stability,
133n
“A Statistical Illusion in Judging
Keynesian Models,” 183

*A Theory of the Consumption
Function*, 77n, 96n
Capitalism and Freedom, 184–85
“Choice, Chance, and the
Personal Distribution of
Income,” 184
“The Demand for Money: Some
Theoretical and Empirical
Results,” 121n
*Income from Independent Professional
Practice*, 53n, 182–83
“Money and Business Cycles,”
133n
*The Relationship of Prices to
Economic Stability and Growth*,
133n
“The Role of Government in
Education,” 185
The Supply of Money and
Changes in Prices and Output,
133n
“The Use of Ranks to Avoid the
Assumption of Normality
Implicit in the Analysis of
Variance,” 183
“The Utility Analysis of Choices
Involving Risk,” 184
Theory of the Consumption Function,
183
Friedman, Rose Director, 1n, 146n
Frisch, Ragnar, 69

G

gold reserves and money supply, 170,
172
gold-silver price ratio, 155, 159
Goldsmith, Raymond, 77n, 117n
Gordon, Robert, 1n
government intervention, effect on
unemployment, 16–18, 20

H

Hamburger, M. J., 122n
Harberger, Arnold, 1n, 14
Hicks, J. R., 128, 128n
Higgs, Henry, 30n
Hough, Harold, 146n
human capital, 5, 9, 182

I

income

- cyclical movements in, 93, 104
- flat tax rate, 185
- from property, 32
- rate of change in, 127

indifference-curve analysis hypothesis

- implication of, 34n, 47–54
- utility function and, 42–47, 44, 46

inflation

- in Latin America, 14–15, 18
- as taxation, 159

inflation, government revenue from

- illustrative calculations for, 140–42
- MF formal analysis of, 136–40
- optimum monetary growth and, 143

inflation rates, 11, 12, 13

- maximum revenue from, 141
- and oil prices, 12–13
- and political cohesiveness, 15–16

insurance purchases, 30–33

interest rate and elasticity, 113, 120, 131, 132

investment, 82–85

investing propensities, 124, 129, 129–30n, 132

investment ROI, 30

IS-LM analysis, 128–30, 130n

J

Jaffe and Kleiman, 15, 16

Johnson, H. G., 1n, 120n, 122n, 123, 128n, 132n

K

Kaysen, C., 28, 28n

Keynes, John Maynard, 19, 123

- and interest-elasticity of monetary demand, 131

Tract on Monetary Reform, 131

Keynesian model of consumption function, 72–88

Klein, Benjamin, 140

Knight, F. H., 30n

Kuznets, Simon, 53n, 91, 91n

L

labor forces, changes in, 8

Laidler, David E. W., 122n, 146n

Latin America, inflation pattern in, 14–15, 18

Laughlin, James Lawrence, 146, 152n

Lee, S. Y., 1n

Lindsey, David, 120n

liquidity preference, 120

Logue and Willett, 15

James Lothian, 1n

lottery purchases, 31–33

Lucas, Robert, 1n, 9, 17, 181

M

Macaulay, F. R., 113n

marginal utility hypothesis, 25, 25n, 57, 59n

markets, forward, 185

Markowitz, Harry, 184

Marshak, J., 62n

Marshall, Alfred, 24n, 25n, 26n, 59n

Marty, Alvin L., 136, 136n, 137n

measured velocity, 99–101, 106, 108, 109, 111–12

Meiselman, David, 1n, 16, 140

Meltzer, Allan 1n

Mendershausen, Horst, 54n

Mitchell, Wesley C., 2

monetary theory and policy, motives for, 116–19

money

- as durable good, 97

- nominal quantity of, 94, 127

- supply, and gold reserves, 170, 172

Money and Banking Workshop of the University of Chicago, 1n, 87n, 120n, 135n

Morgenstern, Oskar, 26, 27–28, 34, 36n

Mosteller, F., 58n

Mundell, Robert A., 136, 159

Muth, John, 9

N

natural rate hypothesis, 6–9

Neumann-Morgenstern Utility Index, 59

Nobel, Alfred, 1

Nogee, P., 58n
Nugent, Walter T. K., 150n

O

occupational choices, 29–30
oil crisis, 12
O’Leary, Paul M., 147n

P

Palgrave, R. H. Inglis, 30n
permanent income hypothesis
 implications for, 119
 and income velocity, 96–99
Phelps, E. S., 6
Philips, A. W., 4–5
Phillips Curve, 4–6, 4, 7, 9–10
Pigou effect, 131
Pigou, Arthur C., 131, 131n
political cohesiveness and inflation
 rates, 15–16
productivity and employment, 8–9,
 14

R

rational expectations, 9
Reagan, John H., 146
real wages, 5–6
recession, spending way out of, 10
relative vs. absolute prices, 17
resources
 risk and, 29, 51–52, 56
 use of, 8, 14
Resumption Act, 159
Return on Investment (ROI), 30
Ricardo, David, 151n
risk
 gambling, 30–33
 insurance, 30–33
 stock market and, 30, 32, 41
Rockoff, Hugh, 146n, 158n, 168n
Ross, A. F., 32n

S

St. Petersburg Paradox, 25, 25n, 57,
 59n
Samuelson, Paul, 61–62
Sargent, Thomas, 9
Savage, L. J.

“The Theory of Statistical
 Decision,” 57n
on expected-utility hypothesis, 57–
 71
 saving and investing propensities,
 129, 129–30n, 132
Scheinkman, José, 1n
Schultz, Theodore W., 1n
Schwartz, Anna J., 1n, 89n, 122, 158
Selden, Richard T., 89n
silver, nonmonetary demand for, 174
silver standard
 U.K. historic price levels under,
 163
 U.S. and U.K. hypothetical prices
 under, 166–67
Sjaastad, Larry, 1n, 14
Smith, Adam, 24n, 29–30, 33n
stable rate theory, 19
“stagflation,” 5, 19, 20
standards
 bimetallic, 155, 156–58
 silver, 163, 166–67
Stewart, William M., 146
Stigler, George J., 1n
stock market and risk, 30, 32, 41
Suh, Jim Soo, 135n
Sundqvist, Sven-Ivan, 1n

T

tax evasion, 185
taxation, 159
Teller, Harry M., 147n
Theory of Games and Economic Behavior,
 26
Tintner, Gerhard, 26n
Tolley, George, 120n

U

unemployment
 and government intervention, 16–
 18, 20
 and inflation, 4, 4–5
 “natural rate of,” 8–9, 13,
 and productivity, 8–9
 rates, 11, 12, 13
unions, 185
utility curve, 45f

utility function and indifference-curve analysis, 42–47, 44, 46

V

Valentine, Gloria, 1n

values in social sciences, 2–3

velocity of income

MF tests of explanation, 101–7

nominal and real stocks of money
and, 92–99

permanent income hypothesis
and, 96–99

secular and cyclical behaviors, 90–
92

Vickrey, William, 28n

volatility, effect on unemployment,
16–18

volunteer army, 185

von Neumann and Morgenstern, 34,
34n

on expected utility hypothesis, 57

*Theory of Games and Economic
Behavior*, 63n

von Neumann, John, 26, 27–28, 34,
36n

W

wages, real, 5–6

Weston, George M., 147n

Wicksell, Knut, 8

Willetts and Logue, 15

World War II aftermath, 5, 15

Z

zero interest and elasticity, 132