Economic Quarterly—Volume 94, Number 2—Spring 2008—Pages 97–120

On the Evolution of Income Inequality in the United States

Kevin A. Bryan and Leonardo Martinez

The recent rise in income inequality in the United States has received considerable attention in policy debates.[1] This article discusses individual income inequality trends. In doing so, we summarize results presented in existing work. As in previous studies, the article shows that income inequality has increased since the 1960s—see, for example, Díaz-Giménez et al. (2002), Eckstein and Nagypal (2004), Weinberg and Steelman (2005), and Katz, Autor, and Kearney (2007). Furthermore, our article documents periods characterized by a decline in real income for lower income groups.

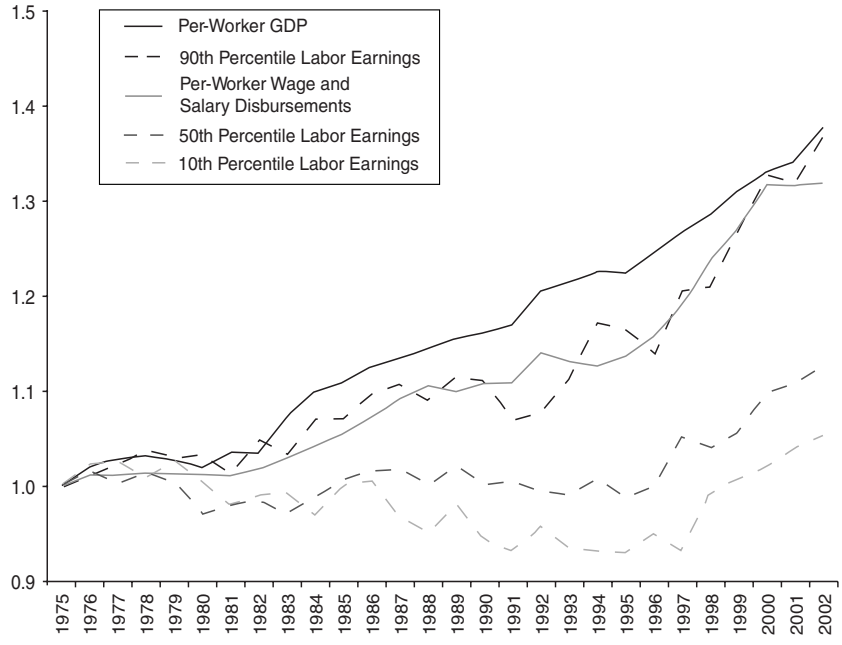
Figure 1 shows that between 1975 and 2002, only labor income in the top 10 percent of the income distribution (Current Population Survey March Supplement) increased more than the per-worker (total nonfarm employment, Bureau of Economic Analysis) wage and salary income (National Income and Product Account).[2] In particular, while during this period per-worker labor income increased 32 percent, labor income in the 10th percentile of the income distribution increased only 5 percent. In addition, Figure 1 shows

The authors would like to thank Kartik Athreya, Andreas Hornstein, Nashat Moin, and Alex Wolman for helpful comments. The views expressed in this article are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mails: Kevin.Bryan@rich.frb.org and Leonardo.Martinez@rich.frb.org.

[1] For instance, it has been discussed recently by George W. Bush, Hillary Clinton, and Ben Bernanke—see Ip and McKinnon (2007), Achenbach (2007), and Bernanke (2007).

[2] Note that in Figure 1, per-worker income and percentile incomes are obtained from different sources. As explained later, the Current Population Survey, our source of percentile incomes, cannot be used to compute total income because income in this survey is topcoded. In order to check whether using different sources is problematic, we also calculated per-worker labor income by using the Current Population Survey to obtain the income for the bottom 90 percent of the distribution and by using the labor income shares of the top 10 percent of the distribution, as computed by Kopczuk, Saez, and Song (2007). We found that the growth of this measure of per-worker income is very similar to the growth of the measure reported in Figure 1.

Figure 1 Real Per-Worker GDP and Earnings (1975 = 1, All Workers)



that between 1975 and 1997, labor income in the 10th percentile decreased 7 percent.

We begin by discussing inequality trends for the whole population, and then we document how these trends vary across different subsets of the population. In doing so, we present findings that are consistent with those in previous studies and are robust to different data sets and inequality measures.

First, we show that the evolution of income inequality displays different patterns for the top and the bottom halves of the income distributions. In the bottom half of the distribution, income inequality rose in the 1980s but was stable after that. Income inequality in the top half of the distribution has risen continuously in recent decades.

Second, we show that trends in male and female income inequality are similar over the past few decades. However, the level of inequality is lower among females than among males. We also show that at the same time inequality among both males and females has been increasing, inequality between the two groups has been decreasing. This decrease in the gender gap implies that overall inequality has been lowered because female incomes caught up with male incomes.

Third, we show that income differentials have increased both between and within levels of education. We also show that the increase in between-education-group inequality has been greater for males than for females.

Our analysis focuses on labor income inequality trends, but brief discussions of wage inequality, welfare inequality, and wealth inequality are also presented. In particular, we discuss why the recent increase in income inequality may not be reflected in an increase in welfare inequality.

Finally, we discuss the pre-1960s period. Although data from before 1960 is fairly limited, studies of wage tables, state censuses, tax returns, and industrial surveys are available. We summarize the findings of these studies, which conclude that U.S. income inequality displayed an inverted U-curve pattern. In the 19th century, income inequality rose, but during the interwar period and especially during World War II, there was a marked decrease in inequality, with narrowing overall income differences, as well as shrinking income gaps between males and females, among different races, among blue- and white-collar workers, and among workers with different levels of education (see, for example, Goldin and Katz 1999a).

The rest of this article is organized as follows. Section 1 describes the data sources we use. Section 2 discusses measures of inequality. Section 3 shows that in recent decades income inequality increased and that this increase in inequality is explained mainly by an increase in inequality among individuals with higher incomes. Section 4 discusses income inequality trends and gender. Section 5 focuses on inequality trends and education. Section 6 comments on wage inequality, welfare inequality, and wealth inequality. Section 7 discusses inequality trends before the 1960s. Section 8 concludes.

1. DATA SOURCES

We use four data sources: the Current Population Survey (CPS) March Supplement, the CPS Outgoing Rotation Group (ORG) supplement, Piketty and Saez's (2003) Internal Revenue Service (IRS) top-income data set, and Kopczuk, Saez, and Song's (2007) Social Security data. The Personal Consumption Expenditures price index is used to deflate income figures—deflating with the CPI-U price index does not materially change our results.

The CPS is a monthly survey of households conducted by the Bureau of the Census. Survey questions are always related to employment, but some months also feature supplemental questions. In particular, the CPS March Supplement (available since 1962, recording income from 1961) asks detailed questions about annual labor income, while the CPS ORG (available since 1979, recording 1978 data) asks about hourly wage and hours worked. Though the CPS collects information on interest payments, social security receipts, and other nonwage income, this data is generally considered less reliable than wage data and as such is often not analyzed in studies of income inequality

(see Luxembourg Income Study 2007). The two CPS supplements are commonly used because of their large sample size (between 60,000 and 190,000 observations) and the length of the sample period.

As is standard when inequality measures are constructed using CPS data, we examine only income from the 10th percentile to the 90th percentile. This is because income data tends to be unreliable at the very bottom of the income distribution, and because CPS data sets are topcoded. That is, incomes above a certain level are capped for privacy reasons. For instance, if an individual earns $200,000 in a year where the cap is $99,999, the CPS would list that individual's income as $99,999. This implies that the CPS offers little guidance for examining the top of the income distribution. This may be a significant problem when analyzing income inequality trends because, as we will show later, over the past decades income inequality has risen very rapidly among the top percentiles of the income distribution and, therefore, using topcoded data biases the measured growth in inequality downward.

For CPS March Supplement data, we use a merged 1962–2003 file compiled by Zvi Eckstein and Eva Nagypal.[3] Our analysis of the CPS ORG data is based on the 2007 National Bureau of Economic Research (NBER) Labor Extracts CD-ROM. Our CPS ORG annual labor income figures are computed by multiplying the NBER ORG Labor Extracts weekly earnings figures by 52. In both CPS files, we keep only full-time, full-year workers, where full-year work is defined as 40+ weeks per year. Volunteers, the self-employed, workers younger than 22 years of age, and workers older than 65 years of age are removed from the sample. As in earlier literature, we multiply topcoded incomes by 1.4. This has little effect since we do not examine top incomes using these data sets, though the topcode is binding for 90th percentile incomes for male college graduates in the mid-1980s. Following Katz, Autor, and Kearney (2007), we drop workers with a stated annualized real wage of less than $1/hr. We drop entries with allocated earnings—meaning that missing data has been imputed—from the CPS ORG. Education dummies are constructed so that 0–11 years of school is "High School Dropout," 12 years is "High School Graduate," 13–15 years is "Some College," 16–17 years is "College Graduate," and 18+ years is "Postgraduate."

Kopczuk, Saez, and Song's (2007) Social Security Earnings Data allows us to study the top percentiles of the income distribution. The authors examine data from individual Social Security returns from 1937 to 2005. Since the data is based on Social Security returns, the income reported only includes pre-tax, pre-transfer wages. In this article, we only analyze publicly available

[3] This file can be found at http://faculty.wcas.northwestern.edu/˜een461/QRproject/.

statistics—income shares—of the Social Security data (which, in general, is not publicly available).[4]

Another data set for high-earner incomes is the one studied by Piketty and Saez (2003) in their examination of income tax returns since 1913. The large number of entries at the top of the distribution in this data set allows us, for instance, to compare the evolution of income of the 99.9th percentile and the 99th percentile of the income distribution. In this article, we analyze summary statistics for labor income made available by Emmanuel Saez.[5] As with the Social Security data, the underlying data set is not publicly available. Labor income data is available from 1927 to 2004 and is missing some years during this period. It should be emphasized that tax data is reported at the level of the tax unit, not the individual. Tax units are sometimes individuals, sometimes couples, and sometimes extended families, depending on how a household chooses to file its taxes and whom it chooses to count as dependents. The increasing correlation between spousal income and compositional changes in tax units makes trends in this data not fully comparable with individual income trends. Because income tax returns are only completed for workers above an exemption limit, it is not possible to examine trends in the bottom of the income distribution with this data set.

2. MEASURES OF INEQUALITY

We measure the degree of income inequality using range ratios and income shares. There are many other commonly used measures of inequality, such as Theil's T, variance of log income, Gini coefficients, the coefficient of variation, and the Atkinson Index. Cowell (1995) provides an overview of benefits and failures of each of these measures.

Range ratios, such as the ratio between the 90th percentile income and the 10th percentile income, are often used because they are easy to understand and unambiguous to compute. Furthermore, they allow us to conduct a quick decomposition of changes in inequality. For instance, we will decompose a change in inequality summarized by a variation in the "90-10 ratio" into changes in the bottom half of the income distribution summarized by a variation in the "50-10 ratio" and changes in the top half summarized by a variation in the "90-50 ratio."

As is standard in studies of income inequality, we focus on logged ratios, because the log of a ratio of two values is equal to the difference of the logs of these values, which is approximately equal to the percentage change between these values. For instance, an increase in the log 90-10 ratio from 0.10

[4] We use summary statistics made available by Wojciech Kopczuk at http://www.columbia.edu/˜wk2110/uncovering/.

[5] See http://elsa.berkeley.edu/˜saez/.

to 0.15 implies that the worker in the 90th percentile went from making approximately 10 percent more than the worker in the 10th percentile to making approximately 15 percent more.

Income shares are simply the share of income held by a given group, such as the top 10 percent of the income distribution. This measure is particularly useful for data sets that do not cover the entire income distribution. For instance, income tax data before World War II covers only the top few percents. Nonetheless, national accounts include total income, and trends in top income shares can therefore be calculated.

3. INEQUALITY TRENDS FOR ALL WORKERS

In this section we focus on pre-tax individual labor income. Focusing on individual income instead of household income allows us to present inequality trends that are not directly affected by changes in household composition. Piketty and Saez (2006) argue that changes in the progressivity of taxes and transfers have been small and, therefore, that pre-tax inequality trends are very similar to after-tax inequality trends.

We study the evolution of inequality since the 1960s. Data availability is significantly better for this period than for earlier periods. Comprehensive micro-level data was only available sporadically before 1940, and decennially from 1940 to 1960. Regular surveys beginning in the early 1960s, such as the CPS March Supplement, offer annual income data along with matched information on education levels, occupations, and other variables. This improved data availability allows us to present a detailed examination of inequality trends.

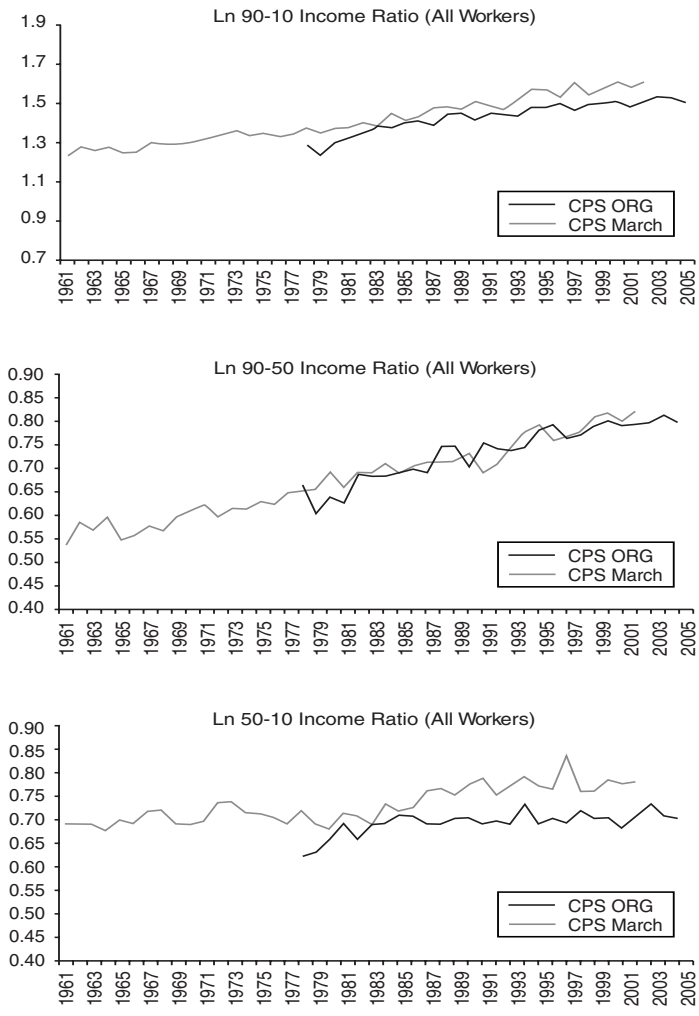
We look at the evolution of the 90-10, 90-50, and 50-10 income ratios. To compute these ratios, we use only the CPS data sets. We do not have exact data for 10th percentile and 50th percentile incomes in the IRS and Social Security data sets used in this article.

Figure 2 presents the evolution of log income ratios. It shows that from 1961 to 2002, the CPS March log 90-10 ratio increased from 1.23 to 1.61. The ratios computed using the CPS ORG data set behave similarly.

Figure 2 also shows that the vast majority of the increase in the log 90-10 ratio is due to an increase in the 90-50 ratio. Since 1961, the log 90-50 ratio grew 0.29, accounting for around 75 percent of the overall increase in 90-10 inequality during this period. The increase in 90-50 inequality also accounts for nearly all of the increase in 90-10 inequality since 1990. This squares with results presented in earlier studies (see, for example, Cutler and Katz 1991 and Katz, Autor, and Kearney 2007). The log 50-10 ratio increased 0.09 during the 1980s but was otherwise constant over the period studied.

The reason for the rise in the 50-10 income ratio during the 1980s has received considerable attention in the income inequality literature. Card and

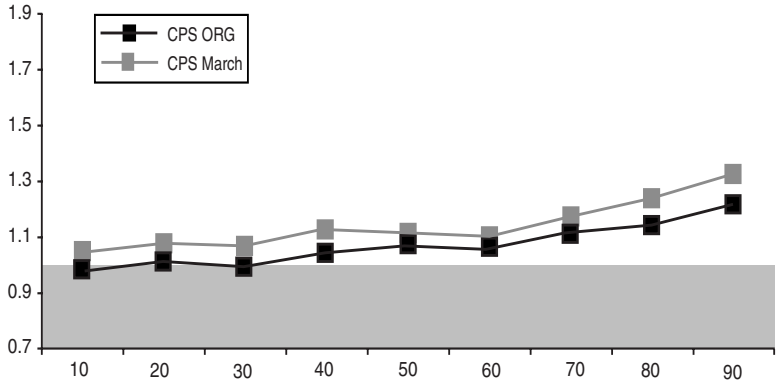
Figure 2 Logged Income Ratios



DiNardo (2002) conclude that the decrease in the real minimum wage is responsible for up to 90 percent of the increase in bottom-half income inequality in the 1980s.[6] Similarly, Lee (1999) uses state-level data on wages and unemployment, and finds that nearly all of the increase in bottom-tail income inequality in the 1980s is a result of changes in the real minimum wage. In

[6] The real minimum wage fell 30 percent between 1980 and 1988. It was roughly stable during the 1990s (Card and DiNardo 2002, Figure 22).

Figure 3 2002–1978 Income Ratios by Percentile for All Workers



contrast, between 1998 and 2006 the real minimum wage fell nearly 20 percent and no significant increase in bottom-half inequality was observed.

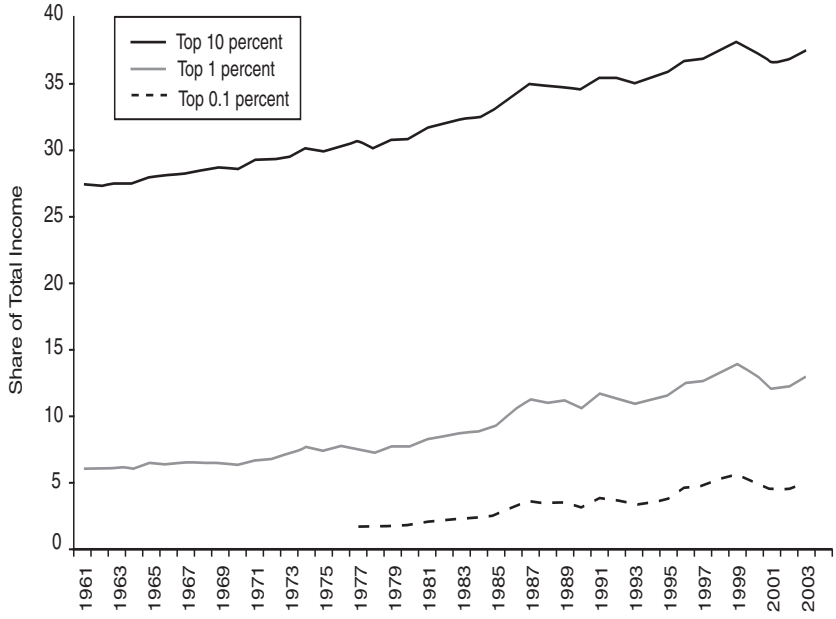
Figure 3 illustrates further that the increase in income inequality during the period under study is concentrated at the top of the income distribution. This figure presents the ratio between the real income in 2002 and the real income in 1978 for each decile of the income distribution. It shows that during this period, differences in income growth rates across percentiles are larger for the higher percentiles.[7] In particular, as in Figure 2, Figure 3 shows that 50-10 inequality increased less than 90-50 inequality during this period.

Since the increase in 90-10 inequality observed in recent decades was concentrated at the top of the 90-10 income distribution, it may also be important to analyze the top 10 percent of the income distribution in order to have a better understanding of the overall trend in inequality. Unfortunately, the CPS data sets are topcoded and therefore do not allow us to conduct such analysis. One way of studying the evolution of income inequality for top incomes is to use Social Security data.

Figure 4 presents the shares of total pre-tax wage earnings of the top 10 percent, the top 1 percent, and the top 0.1 percent of the distribution computed using Social Security data by Kopczuk, Saez, and Song (2007). It shows that between 1961 and 2003, the labor income share of the top 10 percent rose from 27 to 37 percent, and that more than 60 percent of this rise is explained by an

[7] In Figure 3, CPS ORG income growth is lower than CPS March income growth. Although several studies examine differences between CPS ORG data and CPS March data (see, for example, Lemieux 2003, 2006a, and 2006b; Borghans and ter Weel 2004; and Katz, Autor, and Kearney 2007), we are not aware of a comprehensive explanation of the differences between the income growth rates in the two data sets.

Figure 4 Income Share of Top Labor Incomes (Social Security)



increase of the share of the top 1 percent of the income distribution. Kopczuk's data also includes the income share of the top 0.1 percent since 1977. More than 60 percent of the increase of the share of the top percentile between 1977 and 2003 is explained by a rise in the share of the top 0.1 percent. The top 0.1 percent of individuals earn between 2 and 5 percent of the national labor income in our sample.

Though there is much less robust data on working conditions other than labor income, evidence in previous studies suggests that including nonwage income and compensation would increase the growth in inequality observed in recent decades. Pierce (2001) compiles data on fringe compensation from census microdata and finds that including benefits such as leave and health insurance increases the growth of inequality. Mishel, Bernstein, and Allegretto (2006) provide evidence of declining medical insurance and pensions for low-wage workers. Hamermesh (1999) finds that workplace injury rates and the number of nighttime or weekend shifts have fallen more rapidly for high-wage workers than for low-wage workers. These findings suggest that inequality measures based on labor income alone should be taken as a lower bound of the increase in inequality.

4. INEQUALITY TRENDS AND GENDER

In this section we present inequality trends for males and females separately. We will show that trends in male and female income inequality over the past few decades are similar. While in 1961 females represented 34 percent of the labor force, in 2007 they represented 46 percent (Bureau of Labor Statistics).

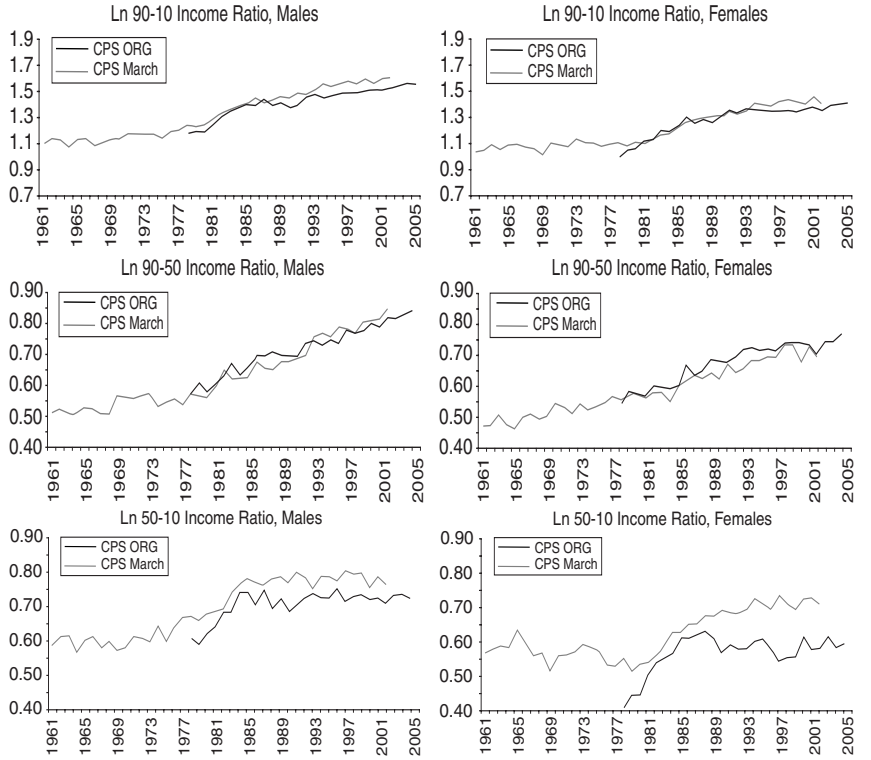
Figure 5 presents the evolution of income ratios for males only and females only. It shows that 90-10 inequality for males has been growing since the late 1960s and that the rate of growth has been higher since the second half of the 1970s. It also shows that 90-10 inequality grew more among males than in the entire population. As in the entire population, the inequality trend for males only is explained by a continuous increase in the 90-50 ratio (which accelerated in the second half of the 1970s) and a rise in the 50-10 ratio concentrated in the 1980s. This is consistent with results presented in previous studies (see, for instance, Katz, Autor, and Kearney 2007).

Figure 5 also shows that the level of inequality is lower among females than among males. The timing of the increase in female inequality is similar to that among males. As in the male population, the increase in inequality among females is mainly explained by an increase in 90-50 inequality and a rise in 50-10 inequality concentrated in the 1980s.

Figure 6 presents the ratios between real incomes in 2002 and 1978 for different percentiles for both males and females (Figure 3 presents the same ratios in the whole population). It shows that the bottom 50 percent of the male income distribution saw no more than a 5 percent increase in real income from 1978 to 2002. The picture is different for females, who have seen rising real wages between 1978 and 2002 across all deciles. Thus, Figure 6 shows that females are driving the income growth at the bottom of the income distribution presented in Figure 3.

While inequality among both males and females has been increasing, inequality between the two groups has been decreasing. Figure 7 presents the evolution of the ratio of female income to male income at the 10th, 50th, and 90th percentiles in the CPS March Supplement data set—the behavior of these ratios in the CPS ORG data set is similar. It shows that, in general, the gender gap is larger at higher levels of income distribution. This is consistent with the fact that inequality is higher among males, as seen in Figure 5. Figure 7 also shows that the gender gap closed substantially over time. The relative increase in female incomes started in the 1970s for the 10th percentile and in the 1980s for the 50th and 90th percentiles. This increase stopped in the mid-1990s. The change in the gender gap implies that overall inequality has been lowered as female incomes caught up with male incomes.

Figure 5 Logged Income Ratios for Males and Females

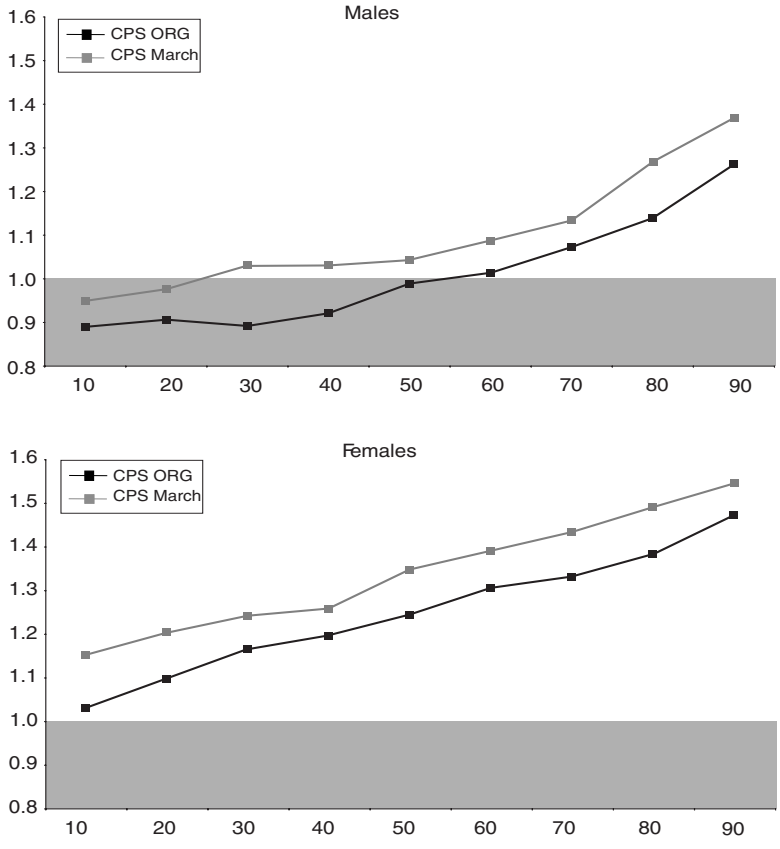


5. INEQUALITY TRENDS AND EDUCATION

In this section we show that inequality has increased both *between* education groups and *within* education groups. That is, real labor income increased more for people with more years of education (an increase in between-group inequality) and the dispersion in labor incomes increased within education groups (within-group inequality increased).

Table 1 presents the evolution of CPS March Supplement male and female labor income for different levels of education. Inequality trends are similar in the CPS ORG data set. This table shows a substantial increase in within-group inequality. For example, for males with a college degree, the 10th percentile income increased 11 percent and the 90th percentile income increased 71 percent between 1963 and 2002. The importance of within-group inequality illustrated in Table 1 is consistent with results in previous studies that show that observable characteristics—mainly education and experience—can only

Figure 6 2002–1978 Income Ratios by Percentile

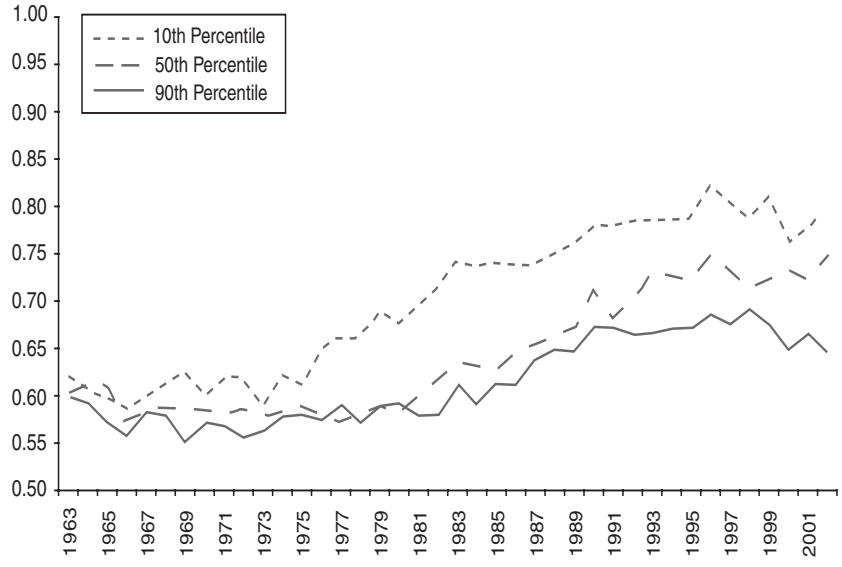


explain a small fraction of observed inequality (see, for example, the discussion in Lemieux 2006b).

An increase in between-group inequality is also present in Table 1. For example, between 1963 and 2002, the median male income increased 78 percent for postgraduates, 41 percent for college graduates, 17 percent for some college, and 11 percent for high school graduates; it decreased 10 percent for high school dropouts. Table 1 also shows that the increase in between-group inequality has been larger for males than for females.

One can also see in Table 1 that there are periods characterized by declines in real income for certain groups. The largest decline is a 27 percent decrease in the median income of high school dropouts between 1972 and 1992. Note that since the 1960s, the percentage of the labor force without a high school degree has halved for both males and females, falling to around 10 percent for

Figure 7 Female-Male Income Ratio



each gender by 2006. The declines in real income seem to have stopped in the 1990s.

A common explanation for the increase in the education premium is skill-biased technological change (SBTC). The SBTC hypothesis suggests that the introduction of computers increased returns to skills, education, and experience, and therefore resulted in a rise in inequality (see, for example, Juhn, Murphy, and Pierce 1993). However, more recent studies challenge this hypothesis by noting that the return to skills grew only in the 1980s and SBTC should have resulted in an increase in the demand for skills in both the 1980s and the 1990s since technological improvements continued into the 1990s (see, for example, Card and DiNardo 2002).

6. WAGE INEQUALITY, WELFARE INEQUALITY, AND WEALTH INEQUALITY

So far, our analysis has focused on annual income inequality trends. In this section we present brief discussions of hourly wage inequality, welfare inequality, and wealth inequality.

Table 1 Real Labor Income (1963=1)

	1972	1982	1992	2002
Postgraduate				
Males 90th Percentile	1.43	1.65	TC	TC
Males 50th Percentile	1.31	1.29	1.44	1.78
Males 10th Percentile	1.40	1.38	1.50	1.64
Females 90th Percentile	1.19	1.25	1.49	1.98
Females 50th Percentile	1.22	1.14	1.33	1.55
Females 10th Percentile	1.22	1.25	1.51	1.74
College Graduate				
Males 90th Percentile	1.34	1.28	1.34	1.71
Males 50th Percentile	1.27	1.15	1.23	1.41
Males 10th Percentile	1.13	1.02	0.95	1.11
Females 90th Percentile	1.14	1.17	1.47	1.86
Females 50th Percentile	1.18	1.15	1.31	1.50
Females 10th Percentile	1.11	1.00	1.09	1.20
Some College				
Males 90th Percentile	1.28	1.20	1.22	1.41
Males 50th Percentile	1.18	1.12	1.06	1.17
Males 10th Percentile	1.15	0.97	0.91	1.04
Females 90th Percentile	1.21	1.32	1.52	1.72
Females 50th Percentile	1.19	1.20	1.33	1.45
Females 10th Percentile	1.15	1.14	1.14	1.23
High School Graduate				
Males 90th Percentile	1.24	1.23	1.20	1.31
Males 50th Percentile	1.25	1.17	1.06	1.11
Males 10th Percentile	1.16	0.95	0.83	0.89
Females 90th Percentile	1.27	1.34	1.45	1.62
Females 50th Percentile	1.18	1.16	1.21	1.33
Females 10th Percentile	1.21	1.18	1.13	1.21
High School Dropout				
Males 90th Percentile	1.31	1.24	1.11	1.14
Males 50th Percentile	1.24	1.07	0.91	0.90
Males 10th Percentile	1.28	1.07	0.88	0.98
Females 90th Percentile	1.19	1.14	1.19	1.25
Females 50th Percentile	1.20	1.15	1.07	1.23
Females 10th Percentile	1.31	1.25	1.15	1.24

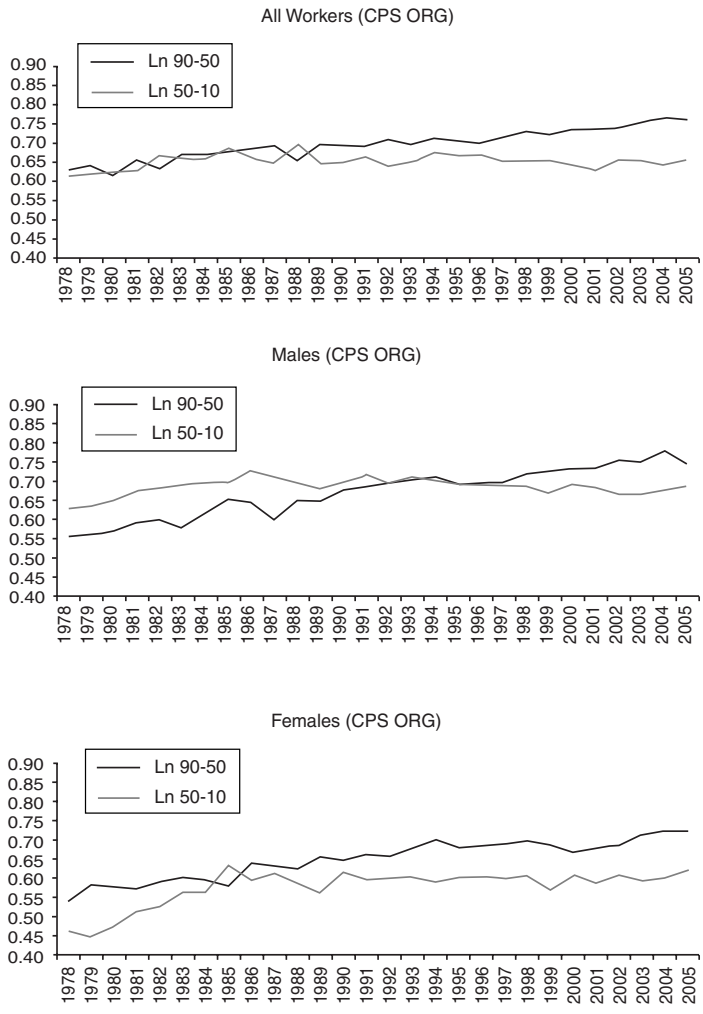
Notes: TC indicates that data was topcoded.

Wage Inequality

Wage inequality trends may be different from the annual income inequality trends discussed in previous sections because of different trends in hours worked across the income distribution.

We construct wage inequality trends using CPS ORG data—as discussed by Lemieux (2006b), CPS March Supplement data only includes intervals of hours worked (e.g., 20–25 hours). The CPS ORG asks hourly workers for

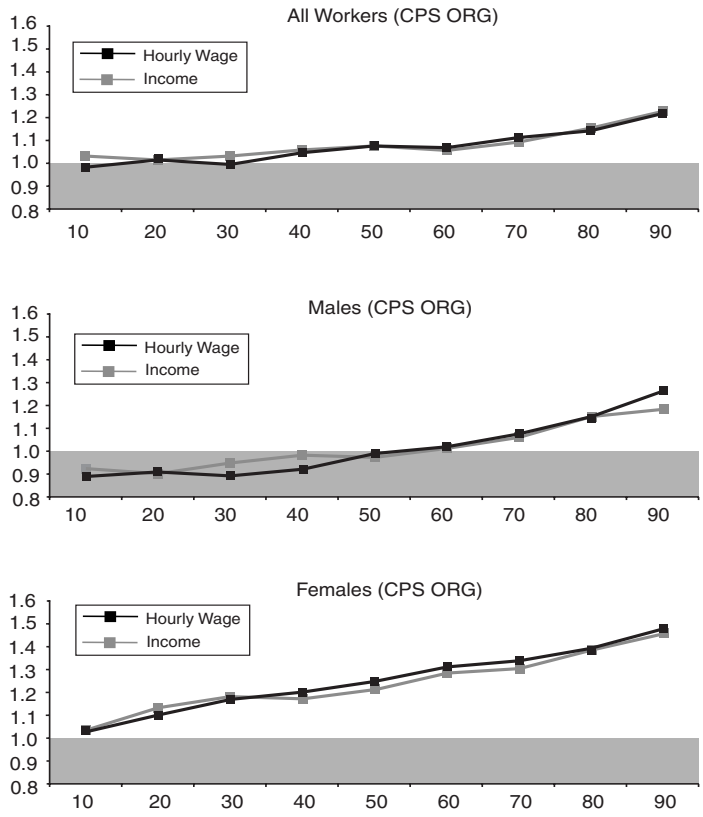
Figure 8 Ln 90-50 and Ln 50-10 Hourly Wage Ratios



their hourly earnings and it asks salaried workers for usual weekly earnings and usual weekly hours worked.

Figure 8 presents logged 90-50 and logged 50-10 wage ratios for all workers, males only, and females only. The figure shows that bottom-tail inequality rose among all groups around the early 1980s, and it increased more among females. Like 90-50 income inequality, 90-50 wage inequality rose continuously from 1978 to 2005. The comparison of Figure 8 with Figures 2 and 5 shows that wage inequality trends are similar to income inequality trends (note that the scale for the horizontal axis in Figure 8 is different than the

Figure 9 2005–1978 Ratios by Percentile



scales in Figures 2 and 5 and, thus, it may appear that inequality increases less in Figure 8 even though this is not the case).

Figure 9 presents the ratio between the real wage in 2005 and the real wage in 1978 for each decile and for all workers, males only, and females only. It also presents the same ratios for real income. The figure shows that the distribution of real wage growth is similar to the distribution of real income growth.

Welfare Inequality

Changes in welfare inequality should not be naïvely inferred from trends in income inequality. Welfare measures depend on the consumption of goods and leisure. It could very well be that while income inequality has increased,

Table 2 Mean Leisure Hours per Week for Males (Aguiar and Hurst 2007)

	Years of Schooling			
Year/Category	**0–11**	**12**	**13–15**	**16+**
1965	104.12	101.66	99.21	101.64
1985	106.94	107.53	105.03	107.02
2003	116.34	108.94	105.42	101.44
Change 1965–2003	12.22	7.28	6.21	-0.20
Change 1985–2003	9.40	1.41	0.39	-5.58

consumption inequality has not increased, or that individuals who benefited from higher consumption growth also experienced a smaller increase in leisure.

Regular surveys on individual consumption have existed since the early 1980s. Krueger and Perri (2006) find both that the level of consumption inequality is lower than the level of income inequality and that consumption inequality increased less than income inequality. They find that, between 1980 and 2003, household income (after-tax labor earnings plus transfers) inequality, measured as the variance of the logs of income in the Panel Study of Income Dynamics (PSID) data set, increased 21 percent.[8] They also find that during the same period, depending on the treatment of durable goods, consumption inequality increased between 2 and 10 percent. Blundell, Pistaferi, and Preston (2006) argue that the difference between the rise in income inequality and the rise in consumption inequality is explained by an increase in the variability of transitory income shocks. They also explain that it is more problematic for low wealth households to insure against these shocks. Attanasio, Battistin, and Ichimura (2004) find a larger increase in consumption inequality than Kreuger and Perri (2006) but nonetheless argue that consumption inequality has increased less than income inequality. These findings indicate that welfare inequality may have increased less than income inequality.

Aguiar and Hurst (2007) examine leisure inequality by aggregating irregular time-use surveys going back to 1965. Leisure is defined as time not spent at work or on household production. They find that the income-poor have seen the largest increase in leisure time. Table 2 shows that, since 1965, leisure has increased the most for those with less education.[9] Since people

[8] Krueger and Perri (2003) find that trends in household income are very similar in equivalent samples of the CPS ORG, the PSID, and the Consumer Expenditure Survey.

[9] This table reports Aguiar and Hurst's (2007) "median" measure of leisure, which includes time sleeping, eating, and activities "pursued solely for direct enjoyment." Note that this definition of leisure does not discriminate between individuals who voluntarily choose not to work and those who are involuntarily unemployed.

with more education have, on average, higher incomes, Aguiar and Hurst's (2007) findings imply relatively larger gains in leisure at the bottom of the income distribution.[10] Thus, these findings also imply that welfare inequality may have increased less than income inequality.

Wealth Inequality

Wealth data is not as readily available as data on income, but surveys such as the Federal Reserve's Survey of Consumer Finances and estate tax returns filings are analyzed in studies of wealth inequality. It is well known that wealth is distributed much more unequally than income. For instance, Casteñada, Díaz-Giménez, and Ríos-Rull (2003) find that in the United States, while the top 1 percent of the wealth distribution holds 26 to 30 percent of the wealth, the income share of the top 1 percent of the income distribution is only 10 to 15 percent of total income.

Trends in income inequality may influence trends in wealth inequality through savings. However, studies have shown that the increase in income inequality observed in recent decades has not been reflected in an increase in wealth inequality. For example, Kopczuk and Saez (2004) find that there has been very little change in the holdings of the top of the wealth distribution since 1970 and that the only major change in the wealth distribution during the 20th century is a massive reduction in the wealth share of the top of the distribution between 1929 and 1945.

7. INEQUALITY TRENDS BEFORE THE 1960S

In this section, we summarize findings of studies of the evolution of income inequality in the United States before the 1960s. There are no large-scale regular population surveys that include individual labor income data during this period. Before 1940, even the decennial U.S. Census did not ask about income (see Williamson and Lindert 1980 and Margo 1999 for discussions of these data limitations). Thus, income inequality before 1940 can only be roughly estimated from sources such as irregular local surveys, state censuses, and tax returns.

Kuznets (1955) famously discusses the basic trends in American income inequality for this period: rising inequality before World War I and falling inequality since the 1920s. Later studies confirmed these trends.

[10] The increase in leisure inequality documented by Aguiar and Hurst (2007) is not inconsistent with the trends in income and wage inequality being similar in Figures 2, 5, 8, and 9. These figures are constructed by considering only full-time workers, and Aguiar and Hurst (2007) construct leisure trends by considering both full-time and part-time workers.

Table 3 Standard Deviation of Manufacturing Wages (Margo 1999, Censuses of Manufacturing)

	1860	1880	Change
Log Wage	0.23	0.36	0.13
Log Wage with State Dummies	0.23	0.32	0.09

There is evidence of increasing wage inequality before the Civil War. For instance, Margo (2000) identifies a compilation of wages paid at government forts for hired labor (clerks, manual laborers, cooks, etc.) from 1820 to 1860. He finds that in this period, wages of clerks rose over a half percentage point more per year than wages of manual laborers. This trend suggests that wage inequality rose—recall that clerks were relatively educated workers in that period. Related wage ratios for skilled artisans and other broad occupation classes show similar patterns. Margo (2000) suggests that this increase in inequality may have been driven in part by a change in the education premium.

Studies also find that income inequality continued to increase, and the premium to skilled labor continued to rise until the end of the 19th century. For example, Table 3 presents the increase in the dispersion of manufacturing wages in the United States from 1860 to 1880 documented by Margo (1999). This increase shows that not only did wage inequality grow across industries, but it also grew within some industries—manufacturing, in this case. Margo (1999) explains that this increase is partially driven by changes in wages across regions after the Civil War. Barro and Sala-i-Martin (1992) report similar trends in their study of the convergence in incomes among states during the postbellum period, documenting a large drop in manufacturing wages in the South. Williamson (2006) provides further evidence of these trends, which he argues are explained in part by the increase in the supply of unskilled labor resulting from high levels of immigration from Europe.

It has also been shown that wage differentials between blue-collar and white-collar workers as well as inter-industry wage differentials shrank around World War I and were stable until the end of the Great Depression. Goldin and Katz (1999a) examine wage series for manufacturing workers, university professors, engineers, and bookkeepers. They find a decrease in the wage premium of the high-education professions over manufacturing wages. Table 4 presents examples of this decrease. The same data show a 20 to 30 percent decrease in the 90-10 wage ratio among manufacturing workers in a number of different industries from 1890 to 1940. Most of this change is concentrated in the bottom half of the distribution. Further, a 1915 Iowa Census was conducted containing information on both income and education, which can then be compared to 1940 United States census data restricted to include only entries in Iowa. Goldin and Katz (1999b) use this data to estimate the return in wages

Table 4 Ratio of Wages of Educated Workers over the Average Manufacturing Wage (Goldin and Katz 1999a)

	Starting Engineers	Male Clerical Workers
1895	—	1.691
1909	1.202	1.652
1914	1.149	1.696
1919	1.005	1.202
1929	1.037	1.128
1939	1.008	1.150
1949	1.012	1.076
1959	—	1.019

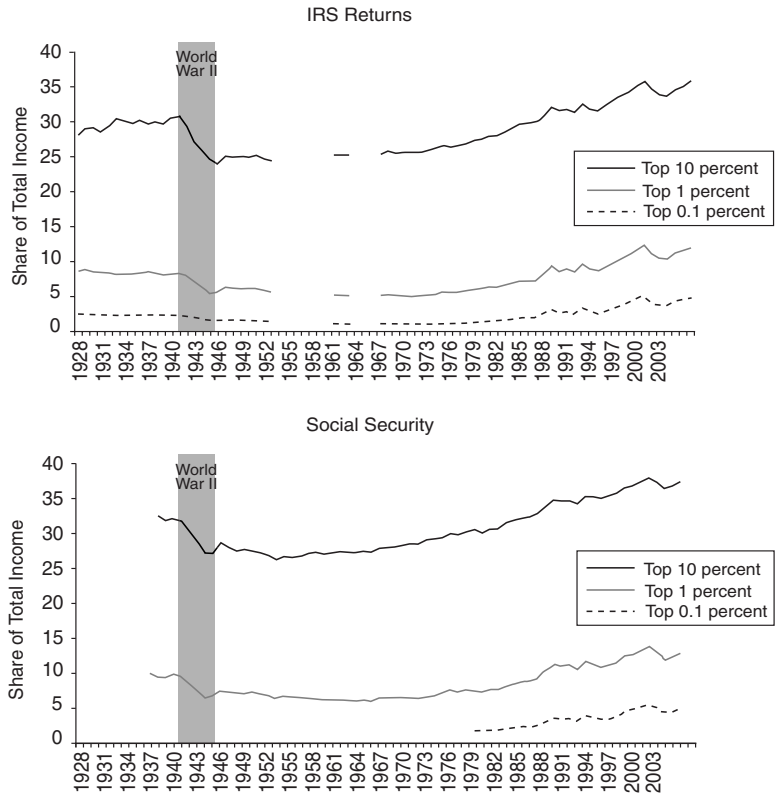
to a year of high school education and find a decrease in this return from 13 percent in 1915 to around 9.5 percent in 1940.

The period around World War II is characterized by decreases in income inequality, an event often called "The Great Compression." Goldin and Margo (1992) explain that this compression is accounted for in part by the National War Labor Board's control of wages during the war. They study public use microdata samples from the 1940 and 1950 censuses and find a large drop in income inequality during this decade, with a low level of income inequality persisting through the 1960s. The return to a year of education computed by Goldin and Katz (1999b) fell two to four percentage points between 1940 and 1950. Piketty and Saez's (2003) data on annual labor income reported in tax returns to the IRS, and Kopczuk, Saez, and Song's (2007) Social Security data show a large drop of the relative income of the top earners around World War II. Figure 10 presents the behavior of the income shares in these two data sets. Although IRS data uses tax units income rather than individual income, the behavior of the two series is quite similar.

8. CONCLUSIONS

This article documents an increase in income inequality in the United States in recent decades. Furthermore, the article documents periods characterized by a decline in real income for lower income groups. We show that this increase in inequality is explained mainly by an increase in inequality at the top of the income distribution. Significant increases in inequality within lower incomes are only observed during the 1980s. We also explain that welfare inequality may have increased less than income inequality. Finally, we show that the recent period of increasing inequality followed a period of decreasing inequality since World War I, which in turn followed a period of increasing inequality in the 19th century.

Figure 10 Income Share of Top Labor Incomes



REFERENCES

Achenbach, Joel. 2007. "Clinton and Inequality." *The Trail*. http://blog.washingtonpost.com/the-trail/2007/10/12/post_134.html.

Aguiar, Mark, and Erik Hurst. 2007. "Measuring Trends in Leisure: The Allocation of Time over Five Decades." *Quarterly Journal of Economics* 122 (August): 969–1006.

Attanasio, Orazio, Erich Battistin, and Hidehiko Ichimura. 2004. "What Really Happened to Consumption Inequality in the US?" Working Paper 10338. Cambridge, Mass.: National Bureau of Economic Research. (March).

Barro, Robert J., and Xavier Sala-i-Martin. 1992. "Convergence." *Journal*

of Political Economy 100 (April): 223–51.

Bernanke, Ben. 2007. "The Level and Distribution of Economic Well-Being." Speech delivered to Greater Omaha Chamber of Commerce, Omaha, February 6.

Blundell, Richard, Luigi Pistaferi, and Ian Preston. 2006. "Consumption Inequality and Partial Insurance." Mimeo, Stanford University.

Borghans, Lex, and Bas ter Weel. 2004. "The Diffusion of Computers and the Distribution of Wages." IZA Discussion Paper No. 1107 (April).

Card, David, and John E. DiNardo. 2002. "Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles." *Journal of Labor Economics* 20 (October): 733–83.

Casteñada, Ana, Javier Díaz-Giménez, and José-Victor Ríos-Rull. 2003. "Accounting for the U.S. Earnings and Wealth Inequality." *Journal of Political Economy* 111 (August): 818–57.

Cowell, Frank. 1995. *Measuring Inequality.* London: Prentice Hall.

Cutler, David, and Lawrence Katz. 1991. "Macroeconomic Performance and the Disadvantaged." *Brookings Papers on Economic Activity* 1991.2: 1–74.

Díaz-Giménez, Javier, Santiago Budria, Vincenzo Quadrini, and Jose-Victor Ríos-Rull. 2002. "Updated Facts on the U.S. Distributions of Earnings, Income and Wealth." Federal Reserve Bank of Minneapolis *Quarterly Review* (Summer): 2–35.

Eckstein, Zvi, and Eva Nagypal. 2004. "The Evolution of U.S. Earnings Inequality: 1961–2002." Federal Reserve Bank of Minneapolis *Quarterly Review* (December): 10–29.

Goldin, Claudia, and Lawrence F. Katz. 1999a. "The Returns to Skill Across the Twentieth Century United States." Working Paper No. 7126. Cambridge, Mass.: National Bureau of Economic Research. (June).

Goldin, Claudia, and Lawrence F. Katz. 1999b. "Human Capital and Social Capital: The Rise of Secondary Schooling in America, 1910 to 1940." *Journal of Interdisciplinary History* 29 (Spring): 683–723.

Goldin, Claudia, and Robert A. Margo. 1992. "The Great Compression: The Wage Structure of the United States at Mid-century." *Quarterly Journal of Economics* 107: 1–34.

Hamermesh, Daniel S. 1999. "Changing Inequality in Markets for Workplace Amenities." *Quarterly Journal of Economics* 114 (November): 1085–123.

Ip, Greg, and John D. McKinnon. 2007. "Bush Reorients Rhetoric, Acknowledges Income Gap." *Wall Street Journal*, March 26, A2.

Juhn, Chinhui, Kevin Murphy, and Brooks Pierce. 1993. "Wage Inequality and the Rise in Returns to Skill." *Journal of Political Economy* 101 (June): 410–42.

Katz, Lawrence, David Autor, and Melissa Kearney. 2007. "Trends in U.S. Wage Inequality: Revising the Revisionists." Working Paper (March).

Kopczuk, Wojciech, and Emmanuel Saez. 2004. "Top Wealth Shares in the United States 1916–2000: Evidence from Estate Tax Returns." Working Paper No. 10399. Cambridge, Mass.: National Bureau of Economic Research. (March).

Kopczuk, Wojciech, Emmanuel Saez, and Jae Song. 2007. "Uncovering the American Dream: Inequality and Mobility in Social Security Earnings Data since 1937." Working Paper No. 13345. Cambridge, Mass.: National Bureau of Economic Research. (August).

Krueger, Dirk, and Fabrizio Perri. 2003. "On the Welfare Consequences of the Increase in Income Inequality in the United States." Working Paper No. 9993. Cambridge, Mass.: National Bureau of Economic Research. (September).

Krueger, Dirk, and Fabrizio Perri. 2006. "Does Income Inequality Lead to Consumption Inequality? Evidence and Theory." *Review of Economic Studies* 73: 163–93.

Kuznets, Simon. 1955. "Economic Growth and Income Inequality." *American Economic Review* 45 (March): 1–28.

Lee, David S. 1999. "Wage Inequality in the United States During the 1980s: Rising Dispersion or Falling Minimum Wage?" *Quarterly Journal of Economics* 114 (August): 977–1023.

Lemieux, Thomas. 2003. "Residual Wage Inequality: A Re-examination." Manuscript, University of British Columbia.

Lemieux, Thomas. 2006a. "Post-secondary Education and Increasing Wage Inequality." *American Economic Review* 96 (May): 195–9.

Lemieux, Thomas. 2006b. "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?" *American Economic Review* 96 (June): 461–98.

Luxembourg Income Study. 2007. *United States 2000: Survey Information*. www.lisproject.org/techdoc/us/us00survey.pdf (October).

Margo, Robert A. 1999. "The History of Wage Inequality in America, 1820 to 1970." Levy Economics Institute Working Paper 286 (November).

Margo, Robert A. 2000. *Wages and Labor Markets in the United States: 1820 to 1860*. Chicago: University of Chicago Press.

Mishel, Lawrence, Jared Bernstein, and Sylvia Allegretto. 2006. *The State of Working America 2006–2007*. Washington, D.C.: Economic Policy Institute.

Pierce, Brooks. 2001. "Compensation Inequality." *Quarterly Journal of Economics* 116 (November): 1493–525.

Piketty, Thomas, and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." *Quarterly Journal of Economics* 118 (February): 1–39.

Piketty, Thomas, and Emmanuel Saez. 2006. "How Progressive is the U.S. Federal Tax System?" Working Paper 12404. Cambridge, Mass.: National Bureau of Economic Research. (August).

Weinberg, John, and Aaron Steelman. 2005. "What's Driving Wage Inequality?" Federal Reserve Bank of Richmond *Economic Quarterly* (Summer): 1–17.

Williamson, Jeffrey G. 2006. "Inequality and Schooling Responses to Globalization Forces: Lessons from History." Working Paper 12553. Cambridge, Mass.: National Bureau of Economic Research. (October).

Williamson, Jeffrey G., and Peter H. Lindert. 1980. *American Inequality: A Macroeconomic History*. New York: Academic Press.

Economic Quarterly—Volume 94, Number 2—Spring 2008—Pages 121–146

On the Sources of Movements in Inflation Expectations: A Few Insights from a VAR Model

Yash P. Mehra and Christopher Herrington

The public's expectations of inflation play an important role in influencing actual inflation and the Federal Reserve's ability to achieve price stability. Hence, there is considerable interest in identifying the economic factors that determine the public's expectations of inflation.[1] In this article, we consider some important macroeconomic determinants of inflation, including commodity and oil prices, and investigate empirically their influences on a survey measure of the public's expectations of inflation from 1953 to 2007, using a structural VAR.[2] We also investigate how the influences of these macroeconomic variables on inflation expectations may have changed during the sample period.

In a recent paper, Leduc, Sill, and Stark (2007) use a structural VAR to investigate the sources of the persistent high inflation of the 1970s. This structural VAR contains a direct survey measure of the public's expectations

The authors would like to thank Kevin Bryan, Robert Hetzel, Pierre Sarte, and John Weinberg for their helpful comments. The views expressed in this article are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] See Ang, Bekaert, and Wei (2006), Bernanke (2007), and Mishkin (2007) for a good introduction to issues related to inflation expectations, actual inflation, and monetary policy. Ang, Bekaert, and Wei provide evidence indicating that survey measures of inflation expectations contain useful information for forecasting inflation. The studies by Bernanke and Mishkin highlight the need for research that promotes a better understanding of the factors that determine inflation expectations and how those expectations affect actual inflation.

[2] Mankiw, Reis, and Wolfers (2003) run single equation regressions relating inflation expectations to several macroeconomic variables. The VAR model, however, allows richer dynamic interactions and, hence, may provide better estimates of the influences of macroeconomic variables on inflation expectations.

of inflation, represented by the median Livingston survey forecast of the eight-month-ahead CPI inflation rate.[3] The other variables in this VAR are actual CPI inflation, a commodity price index, the unemployment rate, a short-term nominal interest rate, and an oil shock variable. The timing of the survey and the way other VAR variables are defined and measured mean the survey participants do not observe contemporary values of VAR variables when making forecasts, thereby helping to identify exogenous movements (shocks) in this survey measure of expected inflation. Leduc, Sill, and Stark (2007) show that the monetary policy response to exogenous movements in expected inflation could explain the persistent high inflation of the 1970s. In particular, prior to 1979 the Federal Reserve accommodated exogenous movements in expected inflation, seen in the result that nominal and real interest rates do not increase in response to such movements, which then led to persistent increases in actual inflation. Such behavior, however, is absent post-1979: The Federal Reserve did not accommodate and aggressively raised nominal and real interest rates, thereby preventing temporary movements in expected inflation from generating persistent increases in actual inflation.[4]

This article uses the structural VAR given in Leduc, Sill, and Stark (2007), denoted hereafter as LSS (2007). While the LSS paper focuses on explaining the sources of the persistently high inflation of the 1970s, this article focuses on explaining the sources of movement in the public's expectations of inflation represented here by the Livingston survey measure of expected inflation. As indicated above, the use of the survey helps identify the exogenous component of expected inflation. We are interested in identifying the role of other macrovariables that may cause movements in expected inflation. Using impulse response functions, we first investigate the responses of expected inflation to temporary surprise movements in macroeconomic variables including expected inflation itself, and using the forecast error variance decomposition of expected inflation, we investigate changes in the relative importance of different macrovariables in explaining the variability of expected inflation.

To investigate how the influences of other macrovariables on expected inflation may have changed over time, we break the whole sample period into one pre-1979 sub-sample, 1953:1–1979:1, and two post-1979 sub-samples,

[3] The participants in this survey are professional forecasters, not the general public. The forecasters are from nonfinancial businesses, investment banking firms, commercial banks, academic institutions, local government, and insurance companies. The survey recently conducted by the Federal Reserve Bank of Philadelphia is biannual. We use this survey primarily because it is the only survey available for the longer sample period covered here. Ang, Bekaert, and Wei (2006) present evidence that indicates the survey contains useful information for predicting future inflation.

[4] The structural VAR contains a short-term nominal interest rate. The behavior of the real interest rate is inferred from the behavior of the nominal interest rate and expected inflation, as the real interest rate is defined as the nominal interest rate minus expected inflation.

1979:2–2001:1 and 1985:1–2007:1.[5] The break in 1979 is suggested by the key result in LSS (2007) that the monetary policy response to exogenous movements in expected inflation changed actual inflation dynamics. It is plausible that monetary policy also changed expected inflation dynamics. The post-1979 sub-sample 1979:1–2001:1 is covered in LSS (2007). We consider another post-1979 sub-sample, 1985:1–2007:1, that we get by modifying the sub-sample 1979:1–2001:1, trimming observations from the initial Volcker disinflation era but including more recent observations from the low inflation period of the 2000s. This sub-sample spans a period of relatively low and stable inflation as its start date corresponds roughly to the beginning of the Great Moderation. The pre-1979 sub-sample includes the period of the Great Inflation of the 1970s.[6] We particularly examine how the influences of different variables on expected inflation may have changed across high and low inflation periods. The use of two post-1979 sub-samples helps us discern the influence of initial Volcker disinflation on post-1979 expected inflation dynamics.

The empirical work presented here suggests several conclusions. First, the survey measure of expected inflation moves intuitively in response to several macroeconomic shocks. Generally speaking, expected inflation increases if there is a temporary unanticipated increase in actual inflation, commodity prices, oil prices, or expected inflation itself, whereas it declines if there is a temporary increase in unemployment. However, the strength and durability of those responses, as well as their relative importance in explaining the variability of expected inflation, have changed considerably over time, especially across pre- and post-1979 sample periods.

Shocks to actual inflation, commodity prices, and expected inflation itself have been three major sources of movement in expected inflation. These three shocks together account for about 95 percent of the variability of expected inflation at a four-year horizon in the pre-1979 sample period, whereas they account for a little over 80 percent of the variability in post-1979 sample periods. The modest decline in the relative importance of these three shocks in explaining the variability of expected inflation is in part due to the decline in the relative contribution of commodity price shocks: They account for about 11 to 22 percent of the variability of expected inflation in post-1979 samples, compared to 40 to 50 percent in the pre-1979 sample period.

---

[5] Other recent research indicating that the responses of inflation to some macroeconomic variables have indeed changed is summarized in Blanchard and Gali (2007) and Mishkin (2007).

[6] Strictly speaking, the first sub-sample period includes the subperiod 1953:1–1965:2 when inflation was also low and stable. Hence, the correct subperiods corresponding to high and low inflation should be 1966:1–1984:1 and 1985:1–2007:1. We, however, follow LSS in breaking up the sample from 1979 for two main reasons. First, the break in 1979 corresponds to the well-known break in the conduct of monetary policy. Second, the use of a somewhat longer sample period (1953:1–1979:1) is necessary for more reliable estimates of VAR parameters, because we have two observations per year due to the use of the Livingston survey data.

Positive shocks to actual inflation, commodity prices, and expected inflation itself lead to increases in expected inflation that are large and long-lasting in the pre-1979 sample period, but muted and short-lived in post-1979 sample periods. The positive response of the real interest rate to several of these shocks, including shocks to expected inflation itself found in the 1979:2–2001:1 sample period but absent in the pre-1979 sample period, is consistent with the view that the above-noted changes in expected inflation dynamics may in part be due to monetary policy, namely, that the Federal Reserve accommodated surprise increases in expected inflation prior to 1979 but not after 1979.

Oil price shocks have only transitory effects on expected and actual inflation in all three sub-sample periods. However, the transitory positive impact of a surprise increase in oil prices on expected inflation has progressively become muted over time, disappearing altogether in the most recent 1985:1–2007:1 sample period. The results also indicate that in response to an unexpected increase in oil prices the real interest rate declines in the pre-1979 sample period, but it increases in post-1979 sample periods. The interest rate responses suggest that the aggressive response of policy to oil shocks since 1979 may in part be responsible for the declining influence of oil prices on expected inflation. The weakened response of inflation expectations to oil price shocks may also explain, in part, the more muted response of actual inflation to oil prices, documented recently in Blanchard and Gali (2007).[7] The result—that there is no longer a significant effect of oil price shocks on inflation expectations—suggests that the Federal Reserve may have earned credibility.

Second, exogenous shocks to expected inflation itself remain a significant source of movement in expected inflation. At a four-year horizon, expectations shocks still account for 35 to 58 percent of the variability of expected inflation in post-1979 sample periods, compared to 36 to 42 percent in the pre-1979 sample period. This result suggests that the Federal Reserve must continue to monitor short-term inflation expectations to ensure that surprise increases in expected inflation do not end up generating persistent increases in actual inflation.

Finally, in the most recent sample period, 1985:1–2007:1, surprise increases in expected inflation die out quickly and expected inflation returns to pre-shock levels within roughly two years after the shock. This response pattern is in the data because the Federal Reserve has not accommodated sudden increases in short-term expected inflation. In such a regime, a positive

[7] Using a VAR, Blanchard and Gali (2007) compare the macroeconomic effects of oil price shocks over two different sample periods, 1970:1–1983:4 and 1984:1–2006:4. Their results also indicate that the response of actual inflation to oil price shocks has become more muted in the more recent sample period. Their VAR, however, does not include inflation expectations and the short-term nominal interest rate and, hence, does not capture the additional channels of expected inflation and policy through which oil prices may affect actual inflation.

shock to short-term expected inflation may lead the public to revise upward their medium- but not necessarily long-horizon expected inflation. Hence, one may find that shocks to short-term expected inflation are no longer correlated with long-term measures of inflation expectations, generating the so-called anchoring of long-term inflation expectations. The fact that one survey measure of long-term inflation expectations—such as the Survey of Professional Forecasters' measure of long-term (10-year) CPI inflation expectations—has held steady since the late 1990s, in contrast to the considerable variation seen before that time, suggests that the public may have come to believe that the Fed would continue not to accommodate temporary shocks to short-term expected inflation.

The rest of the article is organized as follows. Section 1 describes the empirical model. Section 2 presents the empirical results. Section 3 provides further discussion of the results pertaining to expected inflation. Finally, we analyze robustness in Section 4, and provide concluding observations in Section 5.

1. EMPIRICAL METHODOLOGY

Structural Identification

The main advantage of using a structural VAR that contains the Livingston survey measure of expected inflation is that the timing and design of the survey and the way other variables in the VAR are defined and measured help identify exogenous movements in expected inflation. In order to illustrate this identification, consider a VAR that allows for the potential presence of contemporaneous feedbacks among all the five variables contained in the VAR: expected CPI inflation (π_t^e), actual CPI inflation (π_t), the log of a commodity price index (cp_t), the unemployment rate (ur_t), and the three-month Treasury bill rate (sr_t). Shocks to oil prices, captured by disruptions to world oil production due to political events in the Middle East, are assumed exogenous with respect to other variables and therefore are included as a dummy variable (oil_t) in the VAR. We focus on a simple version that allows for only one-period lagged values of endogenous variables as in equation (1):

$$BX_t = \Gamma_0 + \Gamma_1 X_{t-1} + \varepsilon_t, \tag{1}$$

where X is a 5 x 1 vector of variables $\left[\pi_t^e, \pi_t, cp_t, ur_t, sr_t\right]$; B , Γ_0 , and Γ_1 are matrices of structural coefficients; and ε_t is a vector of structural shocks $[\varepsilon_{1t}, \varepsilon_{2t}, \varepsilon_{3t}, \varepsilon_{4t}, \varepsilon_{5t}]$. We assume that structural shocks have zero means and are uncorrelated with each other. B is a 5 x 5 matrix, which contains ones along the main diagonal, and its off-diagonal elements are the structural coefficients that allow for the presence of contemporaneous feedbacks among the variables. We can see this clearly if we explicitly write the equations in the structural VAR, as shown in equations (1.1) through (1.5):

$$\pi_t^e + b_{12}\pi_t + b_{13}cp_t + b_{14}ur_t + b_{15}sr_t = \tag{1.1}$$

$$\tau_{10} + \tau_{11}\pi_{t-1}^e + \tau_{12}\pi_{t-1} + \tau_{13}cp_{t-1} + \tau_{14}ur_{t-1} + \tau_{15}sr_{t-1} + \varepsilon_{1t},$$

$$b_{21}\pi_t^e + \pi_t + b_{23}cp_t + b_{24}ur_t + b_{25}sr_t = \tag{1.2}$$

$$\tau_{20} + \tau_{21}\pi_{t-1}^e + \tau_{22}\pi_{t-1} + \tau_{23}cp_{t-1} + \tau_{24}ur_{t-1} + \tau_{25}sr_{t-1} + \varepsilon_{2t},$$

$$b_{31}\pi_t^e + b_{32}\pi_t + cp_t + b_{34}ur_t + b_{35}sr_t = \tag{1.3}$$

$$\tau_{30} + \tau_{31}\pi_{t-1}^e + \tau_{32}\pi_{t-1} + \tau_{33}cp_{t-1} + \tau_{34}ur_{t-1} + \tau_{35}sr_{t-1} + \varepsilon_{3t},$$

$$b_{41}\pi_t^e + b_{42}\pi_t + b_{43}cp_t + ur_t + b_{45}sr_t = \tag{1.4}$$

$$\tau_{40} + \tau_{41}\pi_{t-1}^e + \tau_{42}\pi_{t-1} + \tau_{43}cp_{t-1} + \tau_{44}ur_{t-1} + \tau_{45}sr_{t-1} + \varepsilon_{4t}, \text{ and}$$

$$b_{51}\pi_t^e + b_{52}\pi_t + b_{53}cp_t + b_{54}ur_t + sr_t = \tag{1.5}$$

$$\tau_{50} + \tau_{51}\pi_{t-1}^e + \tau_{52}\pi_{t-1} + \tau_{53}cp_{t-1} + \tau_{54}ur_{t-1} + \tau_{55}sr_{t-1} + \varepsilon_{5t}.$$

Equation (1.1) relates expected inflation to its own lagged value, current and one-period lagged values of actual inflation, commodity prices, the unemployment rate, and the short-term interest rate, suggesting that expected inflation at time t is likely to be influenced by period t values of other variables in the VAR and, hence, is endogenous. If one is interested in recovering the component of expected inflation that is uncorrelated with contemporaneous (and lagged) values of other VAR variables (namely, the shock ε_{1t}), one needs to impose restrictions on the structural coefficients that allow contemporaneous feedback among variables.

One simple identification strategy used in LSS (2007) assumes expected inflation does not respond to contemporaneous information on actual inflation and the other variables of the VAR. In particular, in this recursive identification scheme we impose the following restrictions on the structural coefficients given in B matrix:

$$\left.\begin{aligned} b_{12} = b_{13} = b_{14} = b_{15} = 0.0 \\ b_{23} = b_{24} = b_{25} = 0.0 \\ b_{34} = b_{35} = 0.0 \\ b_{45} = 0.0 \end{aligned}\right\}. \tag{2}$$

The restrictions given in equation (2) amount to having a B matrix that contains ones along the main diagonal and zeros above, denoting the identification scheme as $\{\pi_t^e, \pi_t, cp_t, ur_t, sr_t\}$. This identification scheme is recursive, meaning a given variable is correlated only with variables that precede it in the

ordering. Thus, the first variable (expected inflation) is not correlated with any other variable of the VAR, the second variable (actual inflation) is contemporaneously correlated only with the preceding expected inflation variable, and so on, and the last variable (short-term nominal interest rate) is correlated with all the preceding variables. This recursive identification scheme is hereafter referred to as benchmark ordering. If we were to focus just on the structural equation for expected inflation, under these restrictions, the expected inflation equation is

$$\pi_t^e = \tau_{10} + \tau_{11}\pi_{t-1}^e + \tau_{12}\pi_{t-1} + \tau_{13}cp_{t-1} + \tau_{14}ur_{t-1} + \tau_{15}sr_{t-1} + \varepsilon_{1t}. \quad (3)$$

Equation (3) is the familiar VAR equation, suggesting that the VAR residuals are estimates of structural shocks to expected inflation under this recursive identification scheme. In general, if we pre-multiply (1) by B^{-1}, we obtain the standard VAR (4):

$$\begin{aligned} X_t &= A_0 + A_1 X_{t-1} + e_t, \quad (4) \\ \text{where } A_0 &= B^{-1}\Gamma_0,\ A_1 = B^{-1}\Gamma_1,\ e_t = B^{-1}\varepsilon_t, \end{aligned}$$

where e_t is a 5 x 1 vector of reduced-form errors, and A_0 and A_1 are matrices of reduced-form coefficients. The identification issue is that of obtaining estimates of structural parameters (B, Γ_0, Γ_1) and structural shocks (ε_t) given estimates of the reduced-form parameters (A_0, A_1) and residuals (e_t). As is well known, we must impose enough identifying restrictions in order to recover structural parameters and shocks. The recursive identification scheme given in (2) imposes 10 restrictions and structural shocks can be recovered using the relationship $\varepsilon_t = Be_t$.[8]

Rationale for Benchmark Ordering

As indicated earlier, the main rationale for the benchmark identification scheme is that the timing and design of the Livingston survey and the way other variables in this structural VAR are defined and measured enable one to assume that the survey participants who forecast CPI inflation at time t do not know the time t realization of inflation and the other variables. Under those assumptions, the restrictions $b_{12} = b_{13} = b_{14} = b_{15} = 0.0$ hold and an expectations shock (ε_{1t}) could be treated as predetermined within the contemporaneous period. As noted previously, the reduced-form error (shock) in the

[8] Quite simply, the identification issue arises because the number of structural parameters we are interested in recovering are usually more than the number of reduced-form parameters that we observe using a reduced-form VAR. Hence, we must impose enough restrictions, thereby reducing the number of structural parameters that need to be recovered. In general, given an $n \times 1$ dimensional VAR and that structural shocks have zero means and are uncorrelated, one needs $\left(n^2 - n\right)/2$ restrictions to identify the structural parameters and shocks. The VAR used here has five variables, so we need 10 restrictions to identify structural parameters and shocks.

expected inflation equation is then an estimate of the structural shock to expected inflation $e_{1t} = \varepsilon_{1t}$.

To analyze robustness we consider an alternative identification ordering. In benchmark ordering, the public's expectations of inflation are not allowed to respond to contemporaneous information on other variables of the VAR, because the public does not observe contemporaneous values of those variables. However, it is plausible that the public has access to other variables that convey information about current values of those variables. Since it is difficult to know what other variables the public may have access to, we examine the sensitivity of our conclusions to an alternative ordering in which expected inflation is ordered last $\{\pi_t, cp_t, ur_t, sr_t, \pi_t^e\}$, thereby allowing expected inflation to respond to contemporaneous information on other variables of the VAR. As indicated later, this alternative ordering yields results that are qualitatively similar to those derived using benchmark ordering.

Measurement of Variables

The structural VAR contains a direct survey measure of the public's expectations of inflation, represented by the median Livingston survey forecast of the eight-month-ahead CPI inflation rate. The participants in this survey are professional forecasters, rather than the general public. Since the Livingston survey is conducted twice a year, the data represent a six-month frequency: May to October and November to April. The timing of the survey and the way the data are measured makes expected inflation a predetermined variable within the contemporaneous period, as explained below.

First, note that survey questionnaires go out to participants in May and November, after the release of the CPI data for April and October, and are returned before the release of the CPI data for May and November. The participants receiving the survey, say, in May (when the CPI for April is known) are asked to predict the level of CPI in December, which is an eight-month forecast. Hence, a forecast of CPI inflation made in period t is measured as the log of the ratio of the expected December CPI level to the actual April CPI level.[9] Other variables of the VAR in period t are then measured as follows: Actual inflation in period t is the log of the ratio of the October CPI level to the April CPI level; the commodity price index, the unemployment rate, and the three-month Treasury bill rate in period t are six-month averages of the monthly data (May to October). Together these observations imply that

[9] The participants receive another questionnaire in November and are asked to predict the level of the CPI in June of the next year, generating a forecast of CPI inflation made in period $t+1$. Actual inflation is for the period between October and April and is constructed as the log of the ratio of the next year's April CPI level to the October CPI level. The CPI, unemployment rate, and the three-month Treasury bill rate in period $t+1$ are six-month averages of the monthly data (November to April).

the survey participants, when making inflation forecasts at time t (namely, in May), do not know the time t realization of actual inflation and other variables in the VAR.

As indicated above, oil price shocks are included as a dummy variable, thereby implicitly assuming they are predetermined. Oil price shocks are measured in two alternative ways. The first method focuses on oil price increases that might be attributed to drops in world oil production due to political events in the Middle East, as in Hamilton (2003). Hamilton identifies the following episodes associated with exogenous declines (in parentheses) in world petroleum supply: November 1956–Suez Crisis (10.1 percent); November 1973–Arab-Israel War (7.8 percent); December 1978–Iranian Revolution (8.9 percent); October 1980–Iran-Iraq War (7.2 percent); and August 1990–Persian Gulf War (8.8 percent). The oil price shock variable is then the oil supply shock variable, included as a quantitative dummy variable that takes a value equal to the drop in world production for these historical episodes, and is otherwise zero.

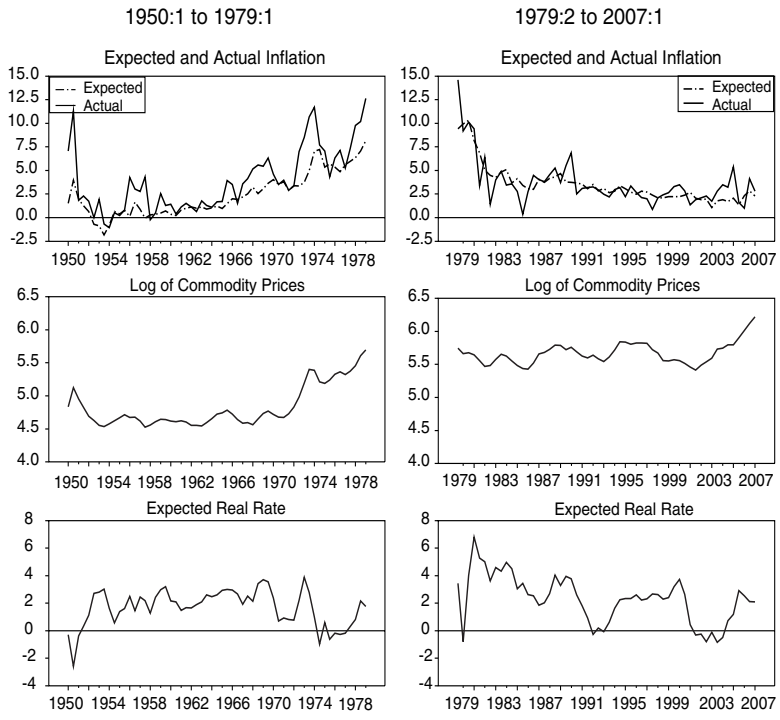
During the most recent period, 1985:1–2007:1, there is only one episode of a drop in world oil production. However, there are several episodes of large increases in oil prices that are due not to drops in world oil production but instead to increases in world demand for oil generated by the growing economies of India, China, and other Asian developing economies. In order to consider such episodes, we consider Hamilton's other measure, net oil price increases, which is a measure of net oil price increases relative to past two-year peaks. We include this measure of net oil price increases as a dummy variable in the VAR, treating it as predetermined with respect to domestic variables included in the VAR. This specification assumes that oil price increases caused by drops in world oil supplies and those caused by increases in world oil demand are alike, having similar consequences for the behavior of macroeconomic variables.[10]

A Visual Look at Data

Figure 1 charts four variables: expected inflation, actual inflation, the log of the commodity price index, and the expected real rate (the three-month Treasury bill rate minus expected inflation). The left panel in Figure 1 charts the data from 1950:1 to 1979:1 and the right panel charts the data from 1979:2 to 2007:1. Several observations stand out. First, even though the actual and expected inflation series move together over time, the Livingston survey participants underpredicted actual inflation when inflation was accelerating and overpredicted inflation during the disinflation of the early 1980s.

[10] Kilian (2007), however, argues otherwise, suggesting it might be important to disentangle the influences of demand- and supply-induced oil price shocks on the economy.

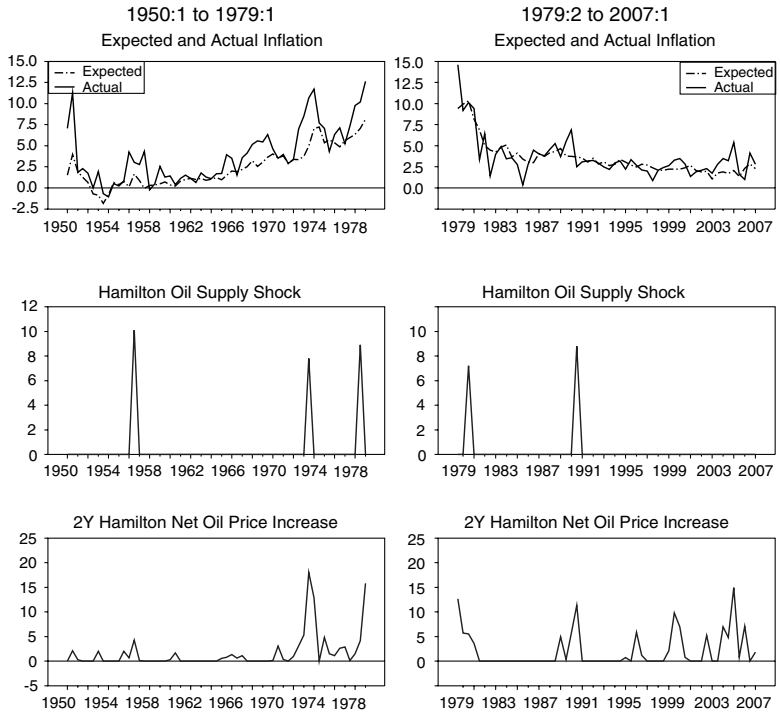
Figure 1 VAR Data



Survey participants could have improved their forecasts by paying attention to actual inflation, suggesting expectations did not respond aggressively to actual inflation. This suggests that the co-movement of the actual and expected inflation series was due more to inflation responding to expectations than expectations responding to inflation. Second, the acceleration in actual inflation does appear to coincide with the pickup in commodity prices. However, the acceleration in inflation appears muted in the post-1985 sample period. Third, Figure 1 also suggests that monetary policy was accommodative in the 1970s. The real interest rate turned negative between 1974 and 1977. By contrast, monetary policy turned very restrictive during the early 1980s, but it again appears accommodative between 2001 and 2004, when the real interest rate turned negative.

Figure 2 charts two measures of oil shocks: one measures drops in world oil production and the other, net oil price increases. Actual and expected inflation are also charted. Two observations stand out. First, oil supply shocks do appear to be associated with spikes in actual inflation in the pre-1979 sample period, but such association appears muted in post-1979 sample periods.

Figure 2 Oil Data



Furthermore, the acceleration in inflation that started during the late 1960s occurred well before the oil shocks of the early 1970s, suggesting that higher oil prices are not a likely explanation of the Great Inflation of the 1970s. Second, in the sample period 1979:2–2007:1, only one episode of a war-related drop in world oil output occurs in 1990, resulting in higher oil prices as measured by net oil price increases. However, the most recent increases in oil prices, as measured by the net oil price increases series, have occurred without a drop in world oil production, suggesting that recent oil price increases could well be due to an increase in global aggregate demand for oil. When comparing the responses of expected inflation to oil shocks across sample periods, the VAR specification employs the second measure of oil shocks, namely, net oil price increases measured relative to past two-year peaks.

Unit Root Properties

As shown in the next section, temporary shocks to some fundamentals (for example, actual inflation, commodity prices) have permanent effects on expected

inflation in the pre-1979 sample period, but not in post-1979 sample periods. But temporary shocks can have a permanent effect on expected inflation only if the latter is a unit root process, suggesting the time series properties of expected inflation must have changed prior to and after 1979. In particular, the expected inflation series must have a unit root in the pre-1979 sample period. This observation is confirmed by the augmented Dickey-Fuller test for unit roots; namely, the test results indicate that both expected and actual inflation series have unit roots in the pre-1979 sample period but are stationary in post-1979 sample periods.[11] In order to identify the fundamentals that may be at the source of generating the permanent changes in expected inflation dynamics, we use a VAR that includes those potential fundamentals other than expected inflation.

2. EMPIRICAL RESULTS

In this section, we examine the responses of expected inflation to different shocks. We focus on shocks to actual inflation, commodity prices, and expected inflation itself, because these three shocks together, as discussed below, account for most of the variability in expected inflation. We also discuss the effects of oil shocks on expected inflation.

Responses of Expected Inflation to Different Shocks

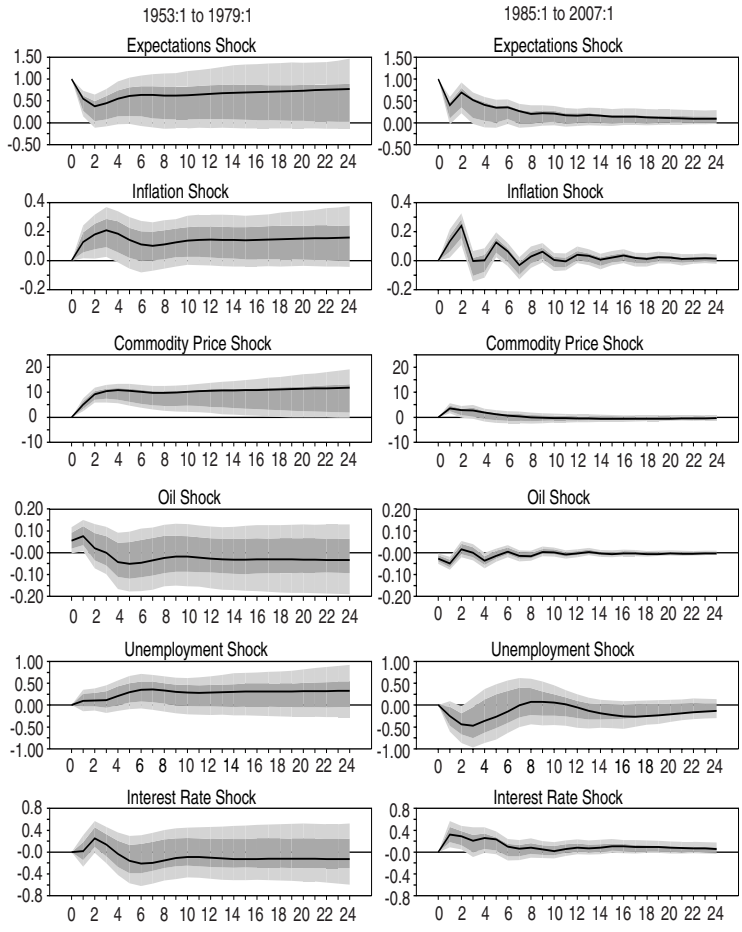
Figures 3 and 4 show the effects of individual, one-time surprise increases in actual inflation, expected inflation, commodity prices, the unemployment rate, interest rate, and oil prices on expected inflation.[12,13] The left panel in Figure 3 shows responses in the Great Inflation (GI) period 1953:1–1979:1, and the right panel shows responses in the Great Moderation (GM) period 1985:1–2007:1; Figure 4 shows responses in the period 1979:2–2001:1 covered in

[11] The test results in LSS (2007) also indicate that expected and actual inflation series have a unit root in the pre-1979 sample period, but are stationary in the post-1979 sample period 1979:1–2001:1 covered there.

[12] Figure 3: The expected inflation responses were generated from a VAR with expected inflation, actual inflation, a CPI, the unemployment rate, the three-month Treasury bill rate, and the Hamilton oil shock variables. For the 1953:1–1979:1 period, oil shock is the shock to the Hamilton oil supply dummy, and for the 1985:1–2007:1 period, oil shock is the shock to the Hamilton net oil price increases. All responses are in percentage terms. The commodity price shock is 100 percent, whereas all other shocks represent 1 percent increases. In each chart, the darker area represents the 68 percent confidence interval and the lighter area represents the 90 percent confidence interval. The x-axis denotes six-month periods.

[13] Figure 4: The expected inflation responses were generated from a VAR with expected inflation, actual inflation, a CPI, the unemployment rate, the three-month Treasury bill rate, and the Hamilton oil supply shock variable. All responses are in percentage terms. The commodity price shock is 100 percent, whereas all other shocks represent 1 percent increases. In each chart, the darker area represents the 68 percent confidence interval and the lighter area represents the 90 percent confidence interval. The x-axis denotes six-month periods.

Figure 3 Expected Inflation Response to...



LSS (2007). In these figures, and those that follow, the solid line indicates the point estimate, while the shaded areas represent 68 percent (darker) and 90 percent (lighter) confidence bands.[14]

Focusing first on the responses of expected inflation to expectations, actual inflation, and commodity price shocks, and comparing them across GI and GM periods as seen in Figure 3, expected inflation increases in response to surprise

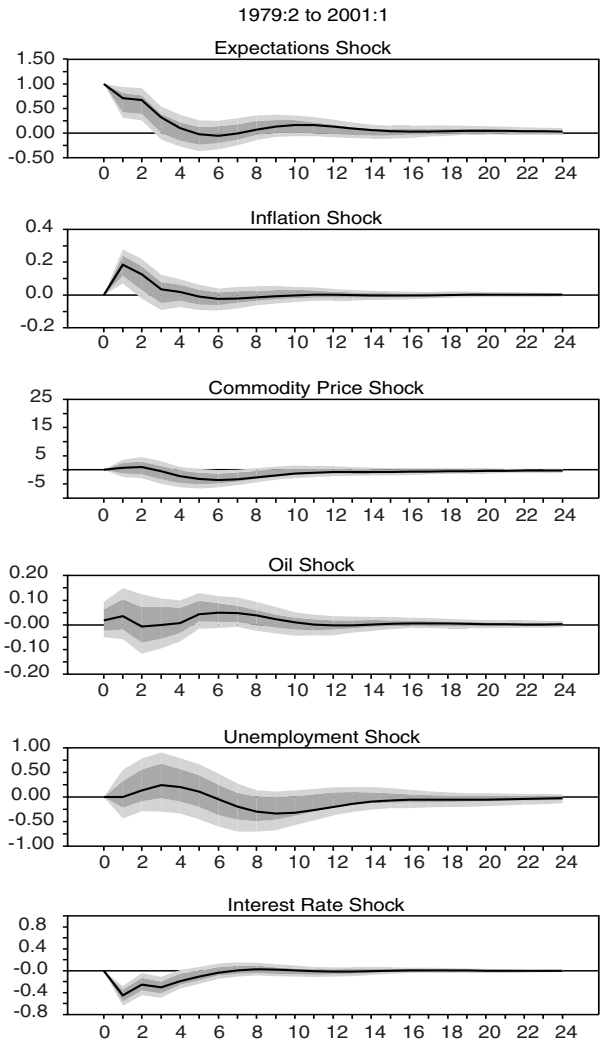
[14] Following LSS (2007), we focus on 68 percent and 90 percent confidence bands. The confidence bands use the bootstrap Monte Carlo method described in Eichenbaum (1998). We would like to thank Keith Sill for providing the programming code used to estimate the confidence bands for the impulse response functions.

increases in each of these three variables. However, both the duration and strength of expectations responses to these three shocks differ substantially across GI and GM sample periods. In the GI period, surprise increases in actual inflation, commodity prices, and expected inflation itself lead to long-lasting increases in expected inflation; in the GM period, those surprise increases have a short-lived effect on expected inflation. To highlight a few features: (a) In response to an expectations shock, expected inflation does not return to its pre-shock level even 12 years after the shock in the GI period, whereas it does so within two years after the shock in the GM period; (b) a similar result holds with respect to the effect of a surprise increase in commodity prices on expected inflation; namely, expected inflation does not return to its pre-shock level in the GI period, whereas it does so within one year in the GM period; (c) in both GI and GM periods, expectations shocks have a much larger effect on the public's expectations inflation than do actual inflation shocks. For example, in the GI period, expected inflation remains at about a .8 percent higher level in response to a one-time 1 percent surprise increase in expected inflation, whereas it remains at about a .2 percent higher level in response to a 1 percent surprise increase in actual inflation. In the GM period, about two years after the shock, expected inflation is still about .4 percent above its pre-shock level in response to a 1 percent surprise increase in expected inflation, whereas it is back to its pre-shock level in response to a 1 percent surprise increase in actual inflation. The previous result also suggests that expected inflation returns more slowly to its pre-shock level after an exogenous shock to expectations than it does in response to an actual inflation shock (see relevant panels in Figure 3).

In traditional Phillips curve inflation models, rising unemployment indicates rising slack in the economy and, hence, should lead the public to expect lower inflation. Similarly, a positive monetary policy shock implies lower inflation and, hence, should lower expected inflation. If we examine the responses of expected inflation to unemployment and monetary policy shocks, the results are mixed (see Figure 3). In response to a surprise increase in the unemployment rate, expected inflation declines only in the GM sample period. The response of expected inflation to a surprise increase in the short nominal interest rate is positive, but these responses are generally not statistically significant. In contrast, the effect of an exogenous oil supply shock on expected inflation is positive and statistically significant in the GI period. However, in the GM sample period, higher oil prices do not have a positive effect on expected inflation. We discuss more about oil price shocks later.

Figure 4 shows the responses of expected inflation to different shocks in the 1979:2–2001:1 sample period. These responses are qualitatively similar to those found in the GM period 1985:1–2007:1 in the sense that shocks lead to changes in expected inflation that are muted and short-lived. Expected inflation still increases in response to a temporary increase in actual inflation or

Figure 4 Expected Inflation Response to...



expected inflation itself. However, a temporary increase in commodity prices, oil prices, or unemployment has no effect on expected inflation. In contrast, expected inflation declines in response to a surprise increase in the short nominal interest rate, and this drop in expected inflation is statistically significant, suggesting monetary policy actions can directly influence the public's expectations of inflation.

Table 1 Variance Decomposition of Expected Inflation

Sample Period 1953:1 to 1979:1

Steps	Ordering: π^e, π, cp, ur, sr					Ordering: π, cp, ur, sr, π^e				
n	π^e	π	cp	ur	sr	π^e	π	cp	ur	sr
1	100.00	0.00	0.00	0.00	0.00	74.06	1.53	2.74	19.32	2.35
2	83.72	4.55	11.39	0.33	0.01	66.04	8.08	8.72	14.86	2.30
3	58.17	8.55	30.82	0.44	2.01	47.53	12.16	25.30	10.03	4.97
4	45.14	11.04	41.55	0.50	1.78	38.75	14.81	34.29	7.56	4.59
8	35.86	8.48	51.34	2.64	1.69	41.62	12.16	40.41	3.77	2.03
16	34.05	7.21	54.26	3.29	1.20	43.88	11.00	41.64	2.50	0.99

Sample Period 1979:2 to 2001:1

Steps	Ordering: π^e, π, cp, ur, sr					Ordering: π, cp, ur, sr, π^e				
n	π^e	π	cp	ur	sr	π^e	π	cp	ur	sr
1	100.00	0.00	0.00	0.00	0.00	58.52	11.16	17.21	10.02	3.09
2	71.42	15.79	0.23	0.00	12.56	38.49	34.19	11.69	5.81	9.82
3	69.51	17.31	0.50	0.25	12.44	36.51	38.15	12.10	4.56	8.68
4	66.96	16.23	0.55	0.97	15.30	37.85	36.35	11.23	4.23	10.34
8	58.07	14.37	10.77	1.82	14.97	35.59	31.91	17.87	4.69	9.94
16	54.47	12.89	13.35	5.89	13.40	33.25	28.72	17.41	11.63	9.00

Sample Period 1985:1 to 2007:1

Steps	Ordering: π^e, π, cp, ur, sr					Ordering: π, cp, ur, sr, π^e				
n	π^e	π	cp	ur	sr	π^e	π	cp	ur	sr
1	100.00	0.00	0.00	0.00	0.00	89.10	7.78	2.83	0.29	0.00
2	74.89	2.69	18.39	3.12	0.91	61.13	10.92	23.41	3.63	0.91
3	58.22	14.05	18.36	3.61	5.76	41.48	25.21	23.37	4.18	5.76
4	52.44	14.34	19.54	5.13	8.56	35.30	25.58	24.77	5.80	8.55
8	54.25	12.93	16.24	6.69	9.89	35.45	25.30	21.58	7.79	9.88
16	50.48	11.08	19.88	9.43	9.13	35.22	22.77	21.99	10.89	9.12

Notes: Entries are in percentage terms with the exception of those under the column labeled "steps." Those entries refer to n-step-ahead forecasts for which decomposition is done.

How important are different shocks in accounting for the variability of expected inflation? Table 1 presents the variance decompositions of expected inflation in three sample periods, with the left panel containing results for benchmark ordering and the right panel for the ordering in which expected inflation is placed last. We focus on the variance of the eight-step-ahead forecast error (which corresponds to four years) that is attributable to each variable of the VAR. As one can see, shocks to actual inflation, commodity prices, and expected inflation itself together account for approximately 95 percent of the variability of expected inflation in the pre-1979 sample period, but account for a little over 80 percent in post-1979 sample periods. The decline in the relative importance of these three shocks that explain variability of expected inflation in post-1979 sample periods is in part due to a decline in the relative contribution of commodity prices: commodity price shocks

account for 11 to 22 percent of the variance of expected inflation compared with 40 to 50 percent in the pre-1979 sample period.

3. MONETARY POLICY EXPLANATION OF THE CHANGE IN THE DYNAMIC RESPONSES OF INFLATION TO SHOCKS

As noted before, Leduc, Sill, and Stark (2007) argue that weakness in the monetary policy response to surprise movements in expected inflation can explain the persistent high inflation of the 1970s. In particular, they find that both nominal and real interest rates rose significantly in response to surprise increases in expected inflation in the post-1979 sample period, but not in the pre-1979 sample period. They interpret this evidence as indicating that the Federal Reserve accommodated increases in the public's expectations of inflation pre-1979, but not post-1979.

Figure 5 reproduces the above-noted result: It charts the dynamic responses of actual inflation, expected inflation, and nominal and real interest rates to an expectations shock, with the graphs in panels A and C covering sample periods 1953:1–1979:1 and 1985:1–2007:1 and the graphs in panel B spanning the sample period 1979:2–2001:1.[15] Note that the real rate increases significantly in response to an expectation shock in the sample period 1979:2–2001:1, whereas such a response is absent in the pre-1979 sample period.[16] In the most recent sample period (1985:1–2007:1) that includes the 2000s, the response of the nominal interest rate to an expectations shock is somewhat muted relative to the 1979:2–2001:1 sample period, so much so that the real rate initially declines and returns to its pre-shock level just one period after the shock (see graphs in panel C).[17] Since this is the sample period during which inflation has been low and stable and inflation expectations stabilized, the interest rate response to a shock to expected inflation is not as aggressive as it was when the Federal Reserve was trying to disinflate during the early 1980s. However, one must be aware of the fact that a shock to expected inflation gets reversed and no longer leads to a persistent increase in actual inflation,

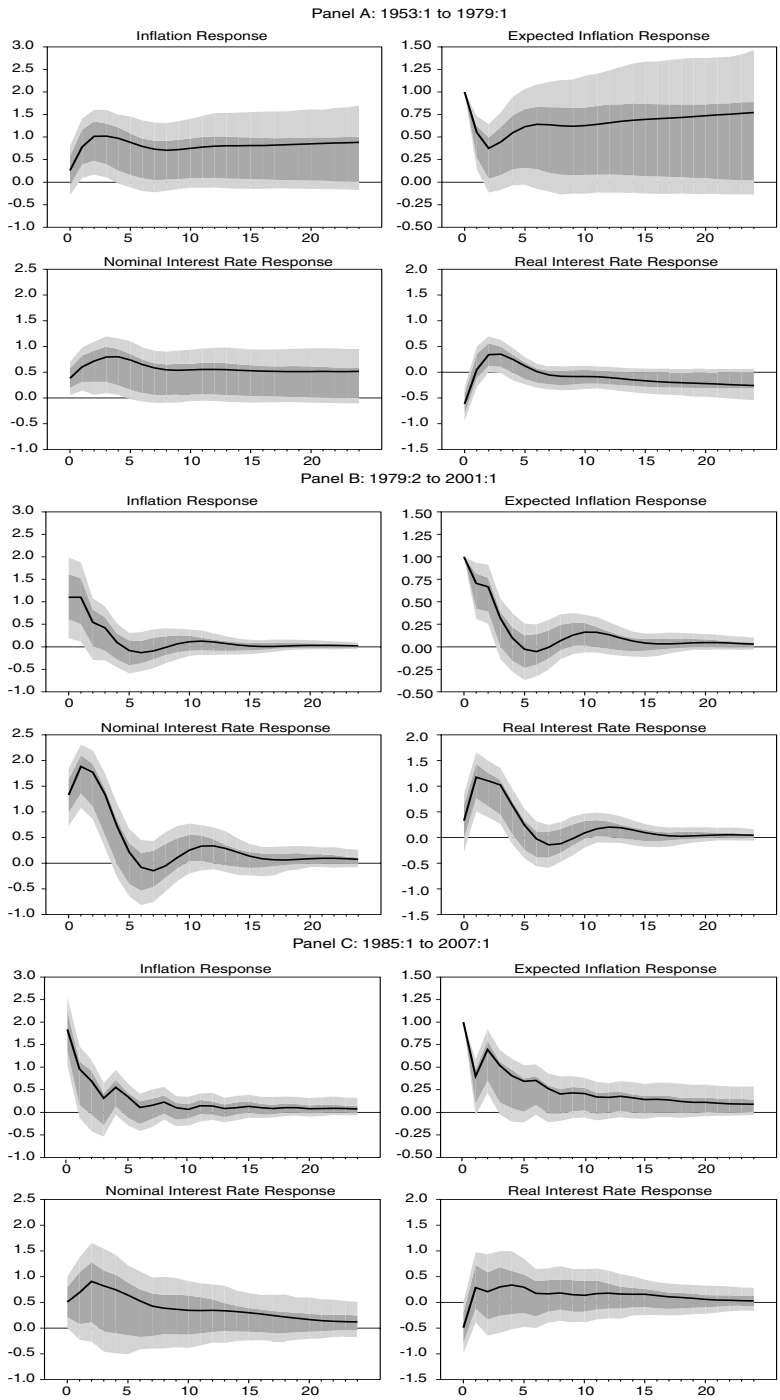
---

[15] Figure 5: Responses to a 1 percent shock to expected inflation. The responses are generated from a VAR with expected inflation, actual inflation, CPI, the unemployment rate, the three-month Treasury bill rate, and a Hamilton oil dummy. For the 1953:1–1979:1 and 1979:2–2001:1 samples, the dummy is the Hamilton oil supply shock. For the 1985:1–2007:1 sample, the dummy is the Hamilton net oil price increase. To conserve space, we report the responses of expected inflation, actual inflation, and nominal and real interest rates. All responses are in percentage terms. In each chart, the darker area represents the 68 percent confidence interval and the lighter area represents the 90 percent confidence interval. The x-axis denotes six-month periods.

[16] As shown in LSS (2007), the strong response of the nominal interest rate to a shock to expected inflation over 1979:2–2001:1 is not driven by the initial Volcker disinflation period. The LSS paper finds such a strong interest rate response over 1982:1–2001:1.

[17] The real interest rate response is constructed as the difference between the nominal interest rate response and the expected inflation response.

Figure 5 Shock to Inflation Expectations



precisely because the public believes the Federal Reserve will continue not to accommodate and, hence, keep inflation low and stable.

Expected Inflation Response to Commodity Prices

As noted above, commodity prices have had significantly less influence on expected inflation over time. The dynamic response of expected inflation to a commodity price shock exhibited in Figures 3 and 4 clearly indicates that the effect of a surprise increase in commodity prices on expected inflation is long lasting in the pre-1979 sample period but short-lived in post-1979 sample periods. Figure 6 shows the responses of nominal and real interest rates to a commodity price shock for three sample periods, in addition to showing the responses of actual and expected inflation.[18] If we focus on the graph for the sample period 1953:1–1979:1, we see that nominal and real interest rates initially increase in response to a surprise increase in commodity prices, but the nominal interest rate does not rise enough to offset the commodity-induced increase in expected inflation, leading to a decline in the real rate. This drop in the real rate persists and is statistically significant, with the expected real rate remaining negative even 12 years after the shock. In contrast, the response of the real interest rate to a commodity shock is quite different in post-1979 sample periods. In particular, in the 1985:1–2007:1 sample period the real interest rate increases and remains positive for about six months after the shock (compare graphs across Panels A and C, Figure 6). These results are consistent with the view that the Federal Reserve's aggressive response to commodity prices is responsible for generating the short-lived response of expected inflation to a commodity shock. The public believes the Fed will continue to restrain inflation, thereby limiting the pass-through of higher commodity prices into expected and actual inflation.

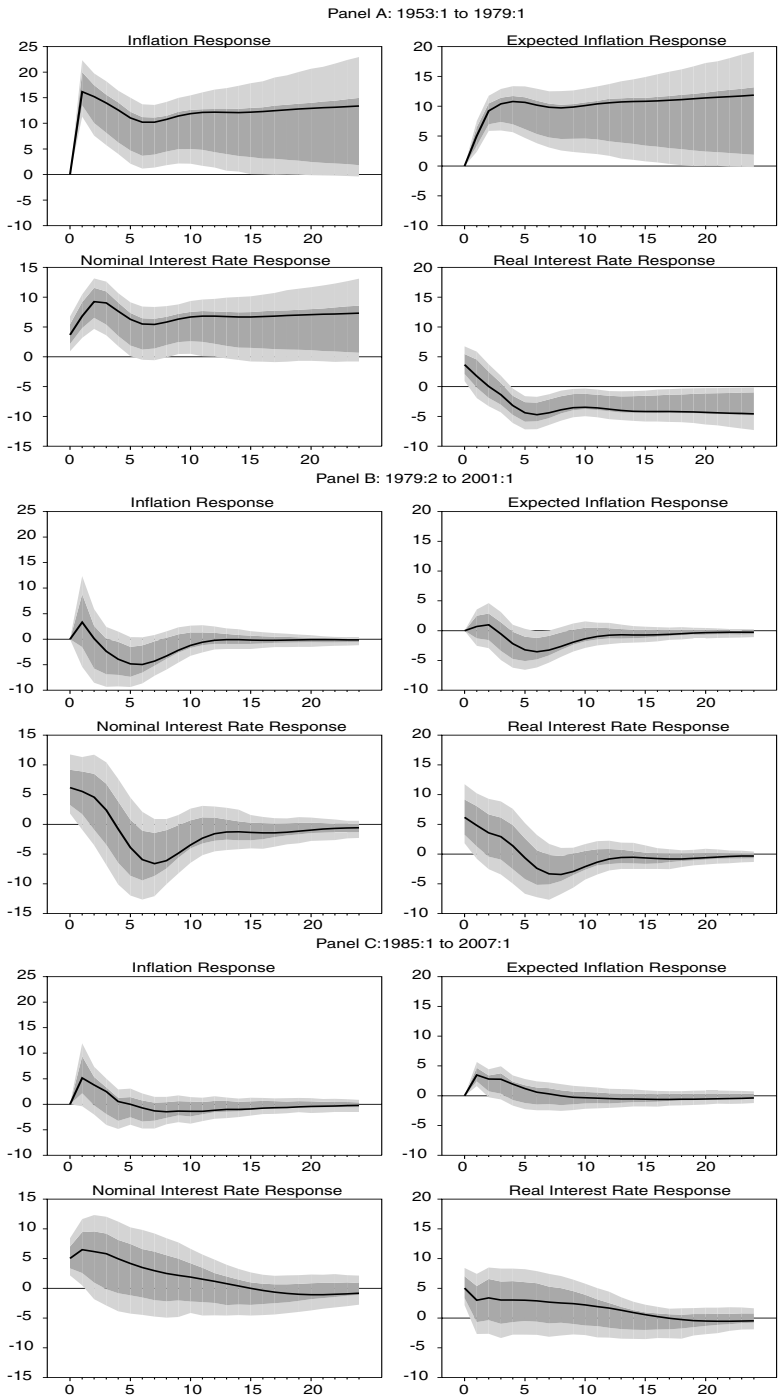
Expected Inflation Response to Oil Price Shocks

Figure 7 shows the responses of actual inflation, expected inflation, nominal interest, and the real interest to oil price shocks.[19] As indicated above,

[18] Figure 6: Responses to a 100 percent shock to the CPI. The responses are generated from a VAR with expected inflation, actual inflation, a CPI, the unemployment rate, the three-month Treasury bill rate, and a Hamilton oil dummy. For the 1953:1–1979:1 and 1979:2–2001:1 samples, the dummy is the Hamilton oil supply shock. For the 1985:1–2007:1 sample, the dummy is the Hamilton net oil price increase. To conserve space, we report the responses of expected inflation, actual inflation, and nominal and real interest rates. All responses are in percentage terms. In each chart, the darker area represents the 68 percent confidence interval and the lighter area represents the 90 percent confidence interval. The x-axis denotes six-month periods.

[19] Figure 7: Responses to a 10 percent shock to the Hamilton net oil price increases. The responses are generated from a VAR with expected inflation, actual inflation, a CPI, the unemployment rate, the three-month Treasury bill rate, and the Hamilton net oil price dummy

Figure 6 Commodity Price Shock



oil price increases that have occurred during the past few years are likely due to increased global demand for oil rather than to disruptions in Middle East oil production. In order to compare the effects of an oil price increase on macroeconomic variables across sample periods, we employ Hamilton's (2003) net oil price increases as the oil shock measure.

The responses to oil price shocks shown in Figure 7 suggest several conclusions. First, oil price shocks have only transitory effects on actual and expected inflation in all three sample periods considered here. Since oil shocks have a transitory effect on actual inflation, it is unlikely that oil shocks can account for the persistently high inflation of the 1970s, as noted in LSS (2007).

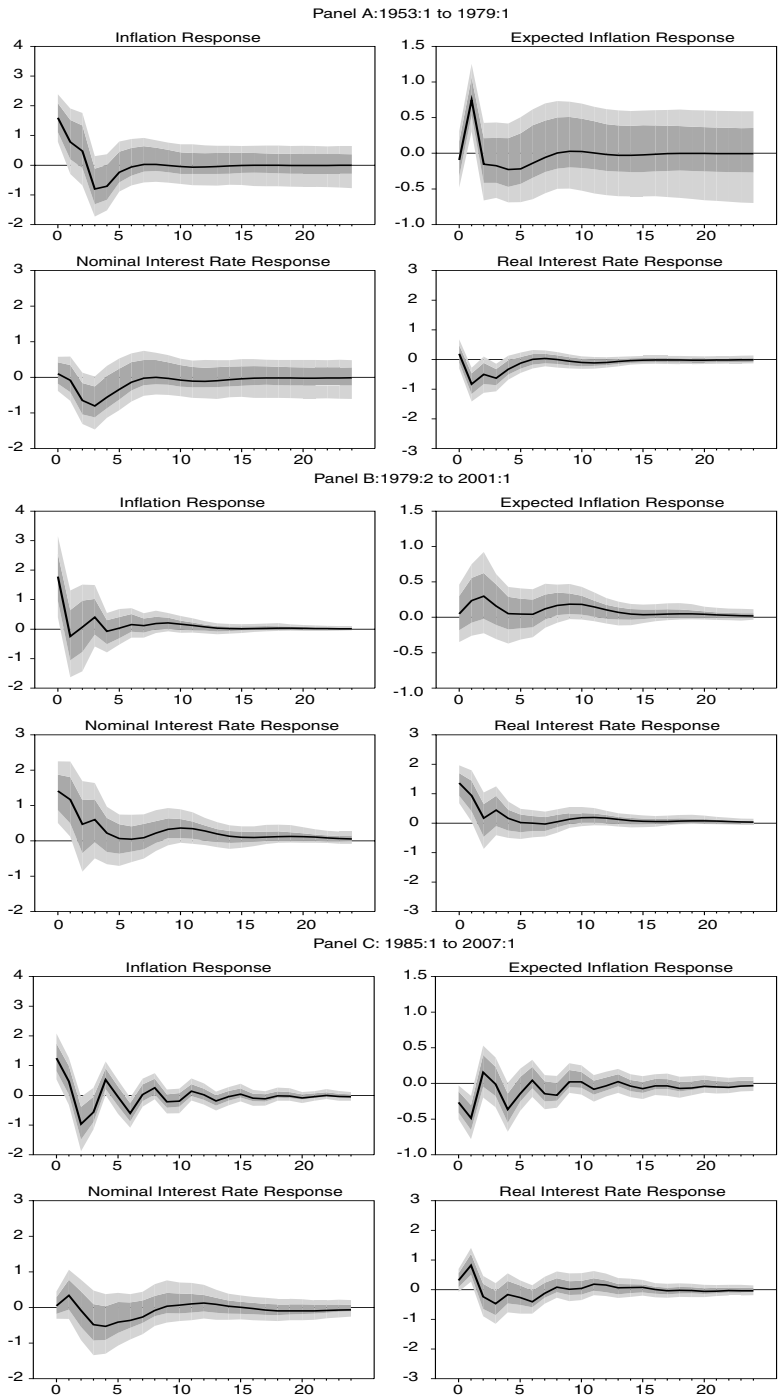
Second, the transitory positive effects of oil price shocks on actual and expected inflation are muted and reversed somewhat sooner in post-1979 sample periods. In the pre-1979 sample period, a positive oil price shock leads a transitory increase in both actual and expected inflation, and those increases are statistically significant (see Figure 7, Panel A). In post-1979 sample periods, however, while a positive oil price shock does lead to an increase in actual inflation, its effect on expected inflation is absent. In fact, in the most recent sample period, 1985:1–2007:1, the initial response of expected inflation to a positive oil price shock is negative and statistically significant. These results appear to be consistent with a view that the public believes the oil-induced increase in actual inflation is likely to be reversed soon and, hence, does not revise its forecast of inflation.

Third, the interest rate responses to oil shocks shown in Figure 7 indicate that monetary policy may in part be responsible for the muted responses of actual inflation to oil shocks found in the most recent sample period, 1985:1–2007:1. In the pre-1979 sample period, the real interest rate declines in response to a positive oil shock, the drop remaining significant up to two years after the shock. In the 1979:2–2001:1 sample period, however, the real interest rate rises significantly following the oil price shock. In the most recent sample period, 1985:1–2007:1, the real interest rate still rises due to a decline in expected inflation. Together these estimates suggest that the aggressive response of policy to oil shocks beginning in 1979 may have been responsible for the muted responses of actual inflation to oil shocks observed in the most recent sample period. The weakened response of expected inflation to oil price shocks may have also contributed to a much more muted response of actual inflation to oil shocks. The negative response of expected inflation to oil shocks also suggests that the public believes the Federal Reserve will continue to restrain inflation and, hence, will not nudge up its forecasts of

---

variables. To conserve space, we report the responses of expected inflation, actual inflation, and nominal and real interest rates. All responses are in percentage terms. In each chart, the darker area represents the 68 percent confidence interval and the lighter area represents the 90 percent confidence interval. The x-axis denotes six-month periods.

Figure 7 Oil Shocks



inflation, despite oil-induced increase in actual inflation. The result—positive oil price shocks do not lead the public to raise its inflation forecast—suggests the Federal Reserve may have earned credibility.

4. EXPECTATIONS SHOCKS: OMITTED FUNDAMENTALS OR SUNSPOTS?

The results pertaining to the variance decomposition of expected inflation presented here indicate that exogenous shocks to expected inflation remain a significant source of movement in expected inflation, even after controlling for its other determinants, such as commodity prices, actual inflation, the unemployment rate, and monetary policy. It is plausible that this VAR does not include some relevant determinants of expected inflation, so that the identified expectations shocks represent the omitted fundamentals. The evidence favoring this view appears in Ang, Bekaert, and Wei (2006), who show that surveys outperform several alternative methods of forecasting inflation and may be capturing information from many different sources not captured by a single model. Moreover, the VAR includes lagged values of fundamentals and, hence, the information captured is backward-looking, whereas survey participants may be responding to information about fundamentals that is forward-looking, namely, the likely expected future values of fundamentals. Finally, the VAR model captures linear relationships among the variables, ignoring any nonlinearity that may be present in the structural equations.

It is equally plausible that exogenous shocks reflect sunspots (nonfundamentals) like random movements in moods of survey participants. In fact, Goodfriend (1993), using a narrative approach, has argued that financial market participants have experienced inflation scares and that, by reacting to inflation scares with a delay, the Federal Reserve generated an upward trend in actual inflation in the 1970s. Such behavior, however, is absent post-1979, when the Federal Reserve, by reacting strongly to inflation scares, prevented such inflation scares from generating persistent increases in actual inflation.

Although it is difficult to identify and test for all potential omitted fundamentals that may be driving movements in expected inflation, the LSS paper does consider some possible candidates. In particular, the paper backs out the structural shocks to expected inflation implied by the VAR model (using the relationship $\varepsilon_t = Be_t$) and then tests whether shocks to expected inflation are related to other macrovariables such as the Producer Price Index, the S&P 500 stock index, the monetary base, and the exchange rates. The results there indicate that none of the variables predict expectations shocks at the 5 percent significance level. But as indicated above, all these variables capture information that is backward-looking. Hence, the issue of whether exogenous movements in expected inflation represent omitted fundamentals or nonfundamentals, akin to inflation scares in Goodfriend (1993), is unsettled.

5. ANALYZING ROBUSTNESS

The major conclusions of this article appear robust to changes in some specifications of the VAR. In particular, in an alternative identification scheme in which we allow expected inflation to respond to all other variables of the VAR within the contemporaneous period, the responses of expected inflation to various shocks do not differ substantially from those found in the benchmark identification, with the exception of the unemployment rate. In particular, expected inflation declines in response to surprise increases in the unemployment rate in both sample periods.

6. CONCLUDING OBSERVATIONS

Using a VAR that includes a survey measure of the public's expectations of inflation represented by the Livingston survey of expected inflation, this article investigates the responses of expected inflation to temporary shocks to several macroeconomic variables over three sample periods, 1953:1–1979:1, 1979:2–2001:1, and 1985:1–2007:1. The empirical work presented suggests that expected inflation moves in an intuitive manner in response to several of these macroeconomic shocks. Generally speaking, expected inflation increases if there is a temporary surprise increase in actual inflation, commodity prices, oil prices, or expected inflation itself, whereas it declines if there is a temporary increase in unemployment. However, the strength and durability of these responses, as well as their relative importance in explaining the variability of expected inflation, have changed considerably across pre- and post-1979 sample periods.

Shocks to actual inflation, commodity prices, and expected inflation itself have been three major sources of movement in expected inflation. These three shocks together account for about 95 percent of the variability of expected inflation at a four-year horizon in the pre-1979 sample period, whereas they account for a little over 80 percent of the variability in post-1979 sample periods. The modest decline in the relative importance of these three shocks in explaining the variability of expected inflation is in part due to the decline in the relative contribution of commodity price shocks: they account for only 11 to 22 percent of the variability of expected inflation in post-1979 sample periods, compared to 40 to 50 percent in the pre-1979 sample period.

The results indicate that temporary positive shocks to actual inflation, commodity prices, and expected inflation itself lead to increases in expected inflation that are long-lasting in the pre-1979 sample period but are muted and short-lived in post-1979 sample periods. This change in the dynamic responses of expected inflation to these shocks across sample periods can be attributed to monetary policy, as the real interest rate rises significantly in response to several of these shocks in post-1985 sample periods, thereby

preventing temporary shocks from generating persistent increases in expected and actual inflation.

The empirical work indicates oil price shocks have only transitory effects on expected and actual inflation in all three sub-sample periods. However, the transitory positive impact of a surprise increase in oil prices on expected inflation has progressively become muted over time, disappearing altogether in the most recent period 1985:1–2007:1. The results also indicate that in response to a surprise increase in oil prices, the real interest rate declines in the pre-1979 sample period, but it increases in post-1979 sample periods. The interest rate responses suggest that the aggressive response of policy to oil shocks since 1979 may in part be responsible for the declining influence of oil prices on expected inflation. The result that there is no longer a significant effect of oil price shocks on inflation expectations suggests the Federal Reserve may have earned credibility.

Exogenous shocks to expected inflation itself remain a significant source of movement in expected inflation. At a four-year horizon, expectations shocks still account for 35 to 58 percent of the variability of inflation expectations in post-1979 sample periods, compared with 36 to 42 percent in the pre-1979 sample. This result suggests that the Federal Reserve must continue to monitor the public's short-term inflation expectations to ensure that surprise increases in expected inflation do not end up generating persistent increases in actual inflation.

Finally, in the recent sample period, 1985:1–2007:1, surprise increases in expected inflation (the measure of short-term inflation expectations) die out quickly, with expected and actual inflation returning to pre-shock levels within about two years after the shock. This response pattern is in the data because the Federal Reserve has not accommodated the increase in actual inflation. In such a regime, a positive shock to short-term expectations may lead the public to revise upward their medium- but not necessarily long-horizon expectations of inflation. Hence, one may find that shocks to short-term inflation expectations are no longer correlated with long-term measures of inflation expectations, generating the so-called anchoring of long-term inflation expectations. The fact that one survey measure of long-term inflation expectations—such as the Survey of Professional Forecasters' measure of long-term (10-year) CPI inflation expectations—has held steady since the late 1990s, in contrast to the considerable variation seen before that time, suggests that the public may have come to believe that the Fed will continue not to accommodate temporary shocks to short-term expectations.

REFERENCES

Ang, Andrew, Geert Bekaert, and Min Wei. 2006. "Do Macro Variables, Asset Markets, or Surveys Forecast Inflation Better?" Finance and Economics Discussion Series 2006-15, Board of Governors of the Federal Reserve System.

Bernanke, Ben S. 2007. "Inflation Expectations and Inflation Forecasting." Remarks at the Monetary Economics Workshop of the NBER Summer Institute, Cambridge, Mass., 10 July.

Blanchard, Olivier J., and Jordi Gali. 2007. "The Macroeconomic Effects of Oil Price Shocks: Why Are the 2000s So Different from the 1970s?" Working Paper 13368. Cambridge, Mass.: National Bureau of Economic Research. (September).

Eichenbaum, M. 1998. "Costly Capital Reallocation and the Effects of Government Spending: A Comment." Carnegie Rochester Conference on Public Policy, 48: 195–209.

Goodfriend, M. 1993. "Interest Rate Policy and the Inflation Scare Problem: 1979–1992." Federal Reserve Bank of Richmond *Economic Quarterly* 79 (Winter): 1–24.

Hamilton, J.D. 2003. "What is an Oil Shock?" *Journal of Econometrics* 113: 363–98.

Kilian, Lutz. 2007. "Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market." University of Michigan and CEPR.

Leduc, Sylvain, Keith Sill, and Tom Stark. 2007. "Self-Fulfilling Expectations and the Inflation of the 1970s: Evidence from the Livingston Survey." *Journal of Monetary Economics* 54: 433–59.

Mankiw, N. Gregory, Ricardo Reis, and Justin Wolfers. 2003. "Disagreement about Inflation Expectations." Working Paper 9796. Cambridge, Mass.: National Bureau of Economic Research. (June).

Mishkin, Frederick S. 2007. "Inflation Dynamics." Working Paper 13147. Cambridge, Mass.: National Bureau of Economic Research. (June).

Economic Quarterly—Volume 94, Number 2—Spring 2008—Pages 147–171

What is the Monetary Standard, Or, How Did the Volcker-Greenspan FOMCs Tame Inflation?

Robert L. Hetzel

What is the monetary standard? Another way to ask this question is to ask how central banks control the price level. In this article, I contrast two views. What I term the "quantity-theory" view implies that to control inflation (with the interest rate as its policy instrument) the central bank needs a policy (reaction function) that relinquishes control of real variables to the price system and that controls trend inflation through the way it shapes the expectational environment in which price setters operate. With credibility, a central bank can allow drift in the price level arising from inflation shocks because these shocks do not propagate. What I term the "nonmonetary" view implies that to control inflation the central bank needs a reaction function whose central element is the manipulation of the difference between the unemployment rate and a full employment benchmark for unemployment subject to the constraint imposed by the Phillips curve. The Phillips curve gives the cost in terms of excess unemployment of preventing inflation shocks from propagating into inflation.

Section 1 exposits the quantity-theory view while Section 2 makes it relevant to actual central bank procedures. Section 3 presents the nonmonetary view. Section 4 treats the contrast between the pre- and post-Volcker periods as an "experiment" in policy procedures useful for choosing between these two views.

The ideas expressed in this article are those of the author and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System. The author gratefully acknowledges helpful criticism from Christopher Herrington, Andreas Hornstein, Thomas Humphrey, Thomas Lubik, Bennett McCallum, and Alexander Wolman. Author e-mail: robert.hetzel@rich.frb.org.

1. THE QUANTITY-THEORY VIEW OF INFLATION

The nominal-real distinction is at the heart of the quantity theory. It arises from the "rationality postulate." Namely, only real variables (physical quantities and relative prices) as opposed to nominal variables (dollar magnitudes) affect individuals' well-being. Because individuals care only about real variables, the implication follows that central banks must care about (control) a nominal variable to control the price level. Central banks possess a monopoly on the creation of the monetary base (a nominal variable). However, because they use the interest rate as their policy variable, money (the monetary base) is determined by market forces. What nominal variable do they control that allows them to influence the behavior of price setters, who care about only real variables (relative prices)? The following explanation proceeds from the insights incorporated in the Cambridge equation of exchange, to the Wicksellian discussion of money supply determination, to the rational expectations discussion of nominal determinacy with central bank interest rate targeting, and finally to discussion of how central banks influence the behavior of price setters.

Equation (1) shows the Cambridge equation of exchange:

$$m_t \cdot \frac{1}{p_t} = k(r_t) \cdot y_t, \tag{1}$$

with m_t the nominal money stock; p_t the price level; $k(r_t)$ the fraction of its income the public desires to hold in the form of money, which depends on the nominal interest rate, r_t; and y_t real income (Pigou 1917). Equation (1) receives content from the assumption that the central bank can cause nominal money, m_t, to change independently of the public's demand for real money (purchasing power), $k(r_t) \cdot y_t$. In these circumstances, the price level will adjust. As a heuristic illustration of how nominal money can change without a prior change in real money demand, Milton Friedman ([1969]1969) made famous the example of a drop of helicopter money.[1]

This formulation is not generally applicable to historical experience because central banks have only rarely attempted to control money directly through targets for monetary aggregates.[2] Nevertheless, what is captured by the quantity-theory appellation is that changes in the price level function as an equilibrating variable in a way that depends on how the central bank controls money creation. In the case in which it pegs its exchange rate to another currency, the price level varies to cause the real terms of trade to vary to equilibrate the balance of international trade. In the case of floating exchange rates, as highlighted in equation (1), the price level (or the goods

[1] On Friedman's contributions to monetary economics, see Hetzel (2007).

[2] For references to central bank attempts to use money targets, see Rich (1987), Neumann (1997), and Hetzel (2008, chap. 13).

price of money—the inverse of the price level) varies to endow the nominal money stock with the purchasing power desired by the public. In this sense, the price level varies to clear the market for the quantity of money. It is this power to control money creation that provides the central bank with control over the domestic price level. But how does it exercise this power? The answer is not obvious because nominal money is demand-determined (determined by the public) given the use of an interest rate by central banks as their policy instrument.

An answer to this question starts with an understanding of a long tradition associated with the name of Knut Wicksell.[3] It is useful to recapitulate this literature from its earliest quantity theory roots in the mid-18th century through its most recent rational expectations formulation in the mid-1980s. The British economist David Hume introduced the central analytical distinction of the quantity theory—the nominal/real dichotomy. Both he and Adam Smith explained how the increase in money caused by the New World gold discoveries would leave the interest rate on capital unaffected.[4] Among others, the later British economists Henry Thornton, David Ricardo, James Mill, Alfred Marshall, and Arthur Pigou emphasized that the productivity of capital determines the real rate of interest relevant to investors (the "natural" rate).

Writing during the suspension of the gold standard at the time of the Napoleonic Wars, Henry Thornton became the first one to understand a central bank as a creator of fiat (paper) money. Thornton was also the first one to explain changes in the supply of money as a manifestation of the difference between the bank rate and the natural rate (the real rate of interest determined by the productivity of capital) (see Hetzel 1987, 9). If the central bank maintains a rate of interest different from this natural rate of interest, the nominal stock of money would change independently of prior changes in real money demand and the price level would have to adjust. In the 1820s, Thomas Joplin associated bank deposit creation with the excess of demand for investment over saving caused by a rate charged on bank loans below the natural rate earned on capital.

Wicksell offered the most famous statement of how changes in the money stock arise when the interest rate set by banks or the central bank differs

---

[3] The discussion draws on Humphrey (1974, 1983b, and 1990). See also Humphrey and Keleher (1982).

[4] Hume ([1752]1956) wrote in *Political Discourses* (cited in Humphrey 1983b, 13): "Money having chiefly a fictitious value, the greater or less plenty of it is of no consequence. . . .[I]f you lent me so much labour and so many commodities; by receiving five per cent, you always receive proportional labour and commodities."

from the natural interest rate.[5] Wicksell ([1898]1965, 120, 148, and 189) also prescribed a price level rule for setting the interest rate peg:

> [T]here is a certain level of the average rate of interest which is such that the general level of prices has no tendency to move either upwards or downwards. . . .Its magnitude is determined by the current level of the natural capital rate and rises and falls with it. If. . . the average rate of interest is set and maintained *below* this normal level. . . prices will rise and go on rising.
>
> [O]nce the entrepreneurs begin to rely upon this process continuing—as soon, that is to say, as they start reckoning on a future rise in prices—the actual rise will become more and more rapid. In the extreme case in which the expected rise in prices is each time fully discounted, the annual rise in prices will be indefinitely great.
>
> If prices rise, the rate of interest is to be raised; and if prices fall, the rate of interest is to be lowered.

What prevents the "entrepreneurs" cited by Wicksell from looking ahead to the "indefinitely great" rise in prices and initiating an immediate rise in prices that prevents any leverage of the central bank over the bank-rate/natural-rate discrepancy? This issue appears in Friedman's ([1968]1969) restatement of the implication that an arbitrary interest rate peg by the central bank would produce an indefinite rise in the price level. By incorporating Irving Fisher's (1896) distinction between nominal and real interest rates, Friedman pointed out that increases in expected inflation would lower the real interest rate corresponding to the nominal rate peg and would thereby intensify money creation and the rise in inflation. Sargent and Wallace (1975, 250) derived an expression that makes the contemporaneous price level a function of the expected future price level and used it to reformulate the Friedman/Wicksell critique of how an arbitrary interest rate peg leaves the price level unanchored. "The public therefore expects that, ceteris paribus, any increase in p_t [the price level] will be met by an equal increase in m_t [the nominal money stock]. . . .There is then nothing to anchor the expected price level." However, McCallum (1981, 1986) pointed out that a central bank that uses the interest rate as its policy instrument can follow a rule that ties down the public's expectation of a nominal variable (either money or the future price level), thereby rendering the price level determinate.[6]

---

[5] Wicksell's analysis did not incorporate the distinction between the nominal and real interest rate developed by Fisher (1896). Friedman ([1968]1969) first combined this distinction with the Wicksell analysis. For a discussion of the history of the distinction between real and nominal interest rates, see Humphrey (1983a).

[6] Goodfriend (1987) extended the analysis by showing that the central bank's rule need only constrain how the public forms its expectation of the price level in response to shocks. Through its loss function, the central bank must care about jumps in the actual price level (relative to the

The McCallum result permits an understanding of actual central bank procedures for controlling inflation by reconciling the endogeneity of money with price level determinacy. His result rests on the rational expectations hypothesis that the central bank can condition the inflationary expectations of price setters (firms) through consistent behavior. But how, given the rationality postulate that requires that the central bank control something real if it is to influence the behavior of private agents whose welfare depends only on real variables? Because central bank use of an interest rate instrument renders money endogenous, its control over prices does not work off a quantitative target for money and a real-balance effect.[7] It follows that the control of prices must derive from the central bank's ability to control the public's expectation of the value of money. Specifically, the central bank must influence the behavior of firms through its control over this expectation. Its control over inflation must work off the desire of firms to set a relative price (a real variable) when they set the dollar price of their product.

One can think of the changes in dollar prices that firms make as comprising two components. The "relative-price-altering" component originates in a desire to change the relative price of their product. The "relative-price-preserving" component originates in a desire to prevent changes in the price level from altering the relative price that firms desire for their product. This component makes dollar price setting depend on forecasts of the future behavior of the price level.[8] Because of these changes in the price level, firms face a coordination problem. Namely, how do they change their dollar price in tandem with the change in the average dollar prices of other firms? The rational expectations hypothesis is that, with respect to the relative-price-preserving component of changes in dollar prices, firms will coordinate on the systematic part of monetary policy. But why should they look to the central bank rather than to some extraneous variable ("sunspots") in solving this coordination problem? As explained below, the central bank has the ability to "shock" real

---

expected price level) and must care about expected changes in the future price level. A central bank concerned about the "inflationary psychology" of bond markets will naturally possess such concerns. The introduction of a third concern beyond the smoothing of actual and expected changes in the price level, namely, a desire to smooth the interest rate, introduces drift in the price level (relative to trend).

One can understand the Goodfriend/McCallum analysis as an application in the monetary area of the general argument for rules made in Lucas ([1980]1981, 255): "[O]ur ability as economists to predict the responses of agents rests, in situations where expectations about the future matter, on our understanding of the stochastic environment agents believe themselves to be operating in. In practice, this limits the class of policies the consequences of which we can hope to assess in advance to policies generated by fixed, well understood, relatively permanent rules (or functions relating policy actions taken to the state of the economy)."

[7] With nominal money fixed, an increase in the price level reduces real money and real spending through the real-balance effect (Patinkin 1965).

[8] In a world of expected price stability, firms only change dollar prices to change relative prices. The enhanced ability of the dollar to serve as a numeraire (a measure of relative prices) is the basis for arguments that the central bank should make price stability its objective.

economic activity through unanticipated money creation (destruction) if the public's inflationary expectations differ from its objective for trend inflation.

To understand this ability, consider the case where the price level evolves unpredictably.[9] Assume, for illustrative purposes, that each period the central bank chooses a random, unannounced target for the price level. In particular, assume that without announcement the central bank sets this period's target for the price level below last period's target. Although individual firms will notice a fall in the demand for their product, that information does not reveal the new price level target.[10] Imagine now a Walrasian "nominal" auctioneer who calls out price levels successively lower than last period's target. Individual firms coordinate reductions in their dollar prices using the auctioneer's announced price level to preserve their relative prices. The process ends when firms resume selling an amount consistent with their profit-maximizing markup.[11] If the central bank behaves in a way that renders the evolution of the price level predictable, the resulting common expectation of the future price level serves the function of the auctioneer.

The rational expectations logic that price setters form their expectations in a way that conforms to the systematic part of monetary policy is that any predictable sequence of price level targets leaves real variables unaffected (apart from possible changes in real money demand). In contrast, if monetary policy causes the price level to evolve in an unpredictable way, it becomes harder for the individual firm to predict how other firms will change their dollar prices. In the case of unanticipated deflation, the first firm to lower its price sells at a loss by selling too much. The price stickiness that accompanies an unpredictable monetary policy shock results from the cost to firms of changing their dollar prices as part of an uncoordinated tâtonnement process to discover the price level consistent with potential output. Because there is a social externality to lowering an individual dollar price to achieve the required reduction in the price level that the individual firm does not capture, individual firms are slow

---

[9] An historical analogue is the real bills period when the Fed tried to restrain what it considered speculation in commodity and stock markets or the stop-go period when it shifted between attempting to target the unemployment rate and inflation (Hetzel 2008, chaps. 3, 23, 24, and 25). For other countries, central bank attempts of uncertain duration to influence the foreign exchange value of their currencies are an example.

[10] The money stock will fall, but variation in the demand for money obscures the implications of nominal money for the price level target.

[11] The auctioneer is omniscient in that he knows that the reduction in aggregate demand is a nominal phenomenon, not a real one due, say, to a perceived reduction in productivity growth that makes the public feel poorer. He also knows when firms' markups (price over marginal cost) return to their profit-maximizing levels. At that time, he ceases to call out reductions in the price level.

The markup is a real variable. Although monetary contraction leads initially to its expansion (assuming no labor hoarding), ultimately firms collectively change their dollar prices to leave the markup at its profit-maximizing (natural) value. See Goodfriend (2004) and Goodfriend and King (1997).

to lower their dollar prices in response to an unanticipated fall in aggregate nominal demand.[12]

One can now answer the question of how the central bank controls the behavior of firms to achieve a desired trend rate of inflation. The self-interest of firms in getting their relative prices right causes them collectively to coordinate on the predictable behavior of the price level in setting price-preserving dollar prices. Of course, that common coordination presupposes the credibility of monetary policy. If the expectation of inflation in the marketplace diverges from the central bank's inflation target, the central bank must create (destroy) money in a way that shocks the real economy.[13] There is a "stick in the closet," but with credibility, the central bank need never take it out.

2. MONETARY CONTROL WITH AN INTEREST RATE INSTRUMENT

The quantity-theory framework reviewed above guides the search for empirical generalizations summarizing central bank behavior that are capable of explaining when the central bank is successful in controlling inflation.[14] This framework implies the necessity for disciplining the central bank reaction function in two ways. First, the central bank must possess procedures that allow it to set the short-term interest rate in a way that tracks the natural rate of interest (i.e., allows the price system to work). The incessant analysis of the real economy engaged in by central banks implies procedures more complicated than the rule advocated by Wicksell of responding directly to the price

---

[12] As a result, the ability of money to serve as a numeraire diminishes. The coordination necessary to allocate resources among specialized markets requires that the price system convey information about the relative scarcity of resources. The requisite economy of communication depends on the use of money as a numeraire. That is, changes in dollar prices should convey information about changes in the relative scarcity of resources. Unpredictable evolution of the price level lessens the ability of money to serve this function. The price system lacks a mechanism for distinguishing between changes in dollar prices required by changes in the scarcity of money and changes in dollar prices required by changes in the relative scarcity of goods. Because there is no way of coordinating the former changes when the price level evolves unpredictably, the dollar prices set by individual firms no longer provide reliable information about the desirability of expanding or contracting output. There is a conflict between the role of the price level as a numeraire and its role as an equilibrating variable that endows nominal money with the purchasing power desired by the public.

[13] The Lucas (1972) Phillips curve, in which the output gap depends on the difference between actual and expected inflation, captures this idea. However, instead of actual inflation the appropriate measure is inflation consistent with the behavior of the central bank. In response to an unanticipated monetary shock that initially impacts output but not inflation, actual and expected inflation may remain identical although expected inflation differs from policy-consistent inflation.

[14] I attribute the success of monetary policy in the Volcker-Greenspan era to its underlying consistency and to the way that consistency shaped inflationary expectations. However, the relentless exercise by the FOMC of reading how the real economy responds to shocks obscures the rule-like behavior of the central bank imposed by the discipline of maintaining low, constant-trend inflation. In contrast to this view, Blinder and Reis (2005) attribute the success of monetary policy in the Greenspan era to the exercise of ongoing discretion. For a more complete discussion, see Hetzel (2008, chap. 21).

level. Second, there must be something systematic in central bank procedures that ties down the way that the public forms its expectation of the future price level (i.e., provides a nominal anchor).

I characterize the underlying consistency in the procedures that restored near price stability in the Volcker-Greenspan era as lean-against-the-wind (LAW) with credibility (Hetzel 2008, chaps. 13–21). Specifically, the FOMC raised the funds rate in a measured, persistent way in response to sustained increases in the rate of resource utilization (declines in the unemployment rate) subject to the constraint that bond markets believed that such changes would cumulate to whatever extent necessary to maintain trend inflation at a low, unchanged rate. In the event of an inflation scare (a sharp jump in the long-term bond rate), the FOMC raised the funds rate more aggressively (Goodfriend 1993; Hetzel 2008, chaps. 13 and 14). Conversely, the FOMC lowered the funds rate in a measured, persistent way in response to sustained declines in the rate of resource utilization subject to the constraint that bond markets believed that such changes would not cumulate to an extent that would raise trend inflation.

The "persistent" part of the "measured, persistent" changes in the funds rate made in response to sustained changes in the degree of resource utilization captures the search for the (unobserved) natural rate.[15] What is important is that the FOMC does not derive its funds rate target analytically from a real intermediate target like excess unemployment but rather follows a procedure that turns determination of the (real and nominal) funds rate over to the working of the economy. Although the FOMC exercises transitory control over the short-term real rate of interest, it does not control the real interest rate in a sustained way.[16] By extension, neither does it determine other real variables such as the unemployment rate (Hetzel 2005, 2006).

Implementation of these procedures required judgment. Much of the FOMC's wide-ranging review of economic activity involved assessment of whether aggregate-demand shocks (changes in resource utilization rates) were sustained or transitory, with only the former calling for funds rate changes. With respect to the "measured" characterization, on rare occasions, incoming data on the economy changed rapidly from offering mixed signals to offering a strong, consistent signal on the change in resource utilization. On these

[15] The natural rate can be thought of as the real interest rate consistent with complete price flexibility (the absence of monetary nonneutrality). Alternatively, one can think of the natural rate as consistent with the operation of the real business cycle core of the economy (Goodfriend 2007).

[16] This assumption lies in the Wicksellian tradition, referred to in Section 1, which assumes that the natural rate of interest is determined by real factors. For example, Pigou (1927, 251) argued for the determination of the real interest rate by real factors, specifically "by the general conditions of demand and supply of real capital. . . .[T]he Central Bank, despite its apparent autonomy, is in fact merely a medium through which forces wholly external to it work their will. Though. . . in determining the discount rate, the voice is the voice of the bank, the hands are not its hands" (cited in Humphrey 1983b, 19).

occasions, for example at the start of the recessions in year-end 1990 and early 2001, the FOMC moved the funds rate by a larger amount than the typical one-quarter percentage point.[17] What is important is not the period-by-period timing of funds rate changes but rather the overall discipline imposed by the requirement of nominal expectational stability. At times of increasing resource utilization, financial markets must believe that funds rate increases will cumulate to whatever extent necessary to maintain trend inflation unchanged at a low level. At times of decreasing resource utilization, markets must believe that funds rate decreases will be reversed when necessary to maintain trend inflation unchanged.

These LAW-with-credibility procedures condition the behavior of financial markets. In response to real aggregate-demand shocks, markets predict the future path of the funds rate necessary to return output to potential, but they do not have to forecast the impact on output of an expansionary or contractionary monetary policy that would force changes in inflation. The resulting continuous variation in the yield curve in response to incoming information on the economy, in which all the variation in future forward rates is real, reduced fluctuations in real output around trend and produced the period of inflation and output stability known as the Great Moderation.[18] The economic forecasts that determine the shape of the yield curve are subject to error, but the process is continually self-correcting. Persistently signed innovations in incoming economic data cause cumulative movements in the yield curve. Note that policymakers and markets "converse" with each other. Central banks do not make public an expected path for the funds rate, but they freely share information about their own forecasts of the economy. Markets then set the yield curve.

The real world counterpart of the quantity-theory thought experiment of an exogenous change in money occurs when markets misforecast the nature and magnitude of a shock for a significant period of time. Consider underestimation by the markets of the magnitude and persistence of a positive real shock so that initially the yield curve fails to rise to the extent required to return real output to trend. Money increases beyond the amount necessary to keep inflation unchanged and portfolio rebalancing occurs (Goodfriend 2000).[19] That

---

[17] Such information implies that the contemporaneous level of the real funds rate differs significantly from its natural value.

[18] For a discussion of the issue of whether the Great Moderation resulted from better monetary policy or fewer macroeconomic shocks, see Velde (2004).

[19] For example, in the last part of the 1980s, the yen appreciated strongly. Under the assumption that this appreciation would dampen export growth and inflation, the Bank of Japan (Finance Ministry) did not raise the discount rate. Given the credibility of monetary policy for price stability, money (M2) growth rose initially without inflation. Portfolio rebalancing appeared in the form of a rise in equity prices and output growth rose strongly (Hetzel 1999). Another example occurred in fall 1998 and spring 1999. At the time, markets widely expected that the Asia crisis would spread and would create worldwide recession and even deflation. In response, the yield

is, money creation causes portfolio holders to rearrange their asset portfolios by buying fewer liquid assets such as bonds and stocks. The prices of these assets rise and their yield falls. In response to the increase in money, the price level rises but without an increase in trend inflation as long as monetary policy remains credible. Especially because of the difficulty of determining the persistence of a shock, it is inevitable that episodes will occur when real shocks push output away from trend and affect the price level. Nevertheless, what is remarkable is how well monetary policy has worked over the last quarter century.

The quantity-theory framework outlined in Section 1 and the above characterization of the FOMC's reaction function in the Volcker-Greenspan era offer a description of the control of inflation in terms of monetary control. Assume that a central bank possesses credibility for a policy of price stability and that its reaction function allows it to set an interest rate peg equal to the natural rate (the rate consistent with perfectly flexible prices). Under this assumption, the central bank merely accommodates changes in the demand for real money associated with whatever real forces drive growth in the real economy plus random changes in real money demand.[20] These are "price-preserving" changes in money.

To illustrate "price-altering" changes in money, consider the example in which the central bank raises its interest rate peg with a lag in response to a permanent real shock to productivity growth that increases the value of the natural rate (Hetzel 2005). The counterpart of the resulting bank rate/natural rate discrepancy is a demand for a flow of services from the capital stock and a flow of consumption that exceeds the amounts given by a hypothetical real economy with completely flexible prices. The price paid for the utilization of resources today is set too low in terms of resources foregone tomorrow. Corresponding to this excess demand for resources is a flow of credit demanded of banks by the public. With a funds rate left unchanged by the central bank, banks accommodate this additional demand through an increase in their deposits. Maintenance by the central bank of the real interest rate below the natural rate is a form of price fixing that creates an excess supply of money (demand for credit) as the counterpart to goods shortages. The concomitant monetary emissions force portfolio rebalancing and changes in the price level

---

curve fell. In the event, the U.S. stock market rose strongly in 1999 and domestic consumption surged (Hetzel 2008, chaps. 17 and 18).

A transitory rise in output (consumption) relative to expected future output (consumption) restrains the rise in the real interest rate (Hetzel 2005).

[20] Money holders who desire additional real money balances sell debt instruments such as Treasury bills to banks and receive demand deposits in return. The central bank accommodates any increase in required reserves as a consequence of maintaining its interest rate peg. Changes in nominal money demand match changes in real money demand so that the price level need not change.

(Hetzel 2004). These are "price-altering" changes in money because they occur with no prior increase in real money demand.

A policy procedure that disciplines money creation to allow only for price-preserving changes in money imposes two sorts of disciplines (real and nominal) that correspond to the two characteristics of the LAW-with-credibility characterization of the Volcker-Greenspan procedures. The first (the real) discipline entails the LAW characteristic whereby the real funds rate tracks the natural interest rate. As long as the central bank maintains the real interest rate equal to the natural rate, real money grows in line with the real money demand consistent with the hypothetical operation of the economy with complete price flexibility and with real money demand shocks.[21] The second (the nominal) discipline entails credibility for maintenance of an unchanged trend inflation rate despite recurrent real aggregate-demand shocks and inflation shocks. Credibility means firms coordinate the relative-price-preserving changes in their dollar prices on the central bank's inflation target. Expected inflation then equals the central bank's inflation target. This level of expected inflation drives an equal amount of money growth and inflation.

The final component of money demand that adds to money growth arises from an inflation target as opposed to a price level target. This component accommodates transitory inflation shocks (relative price shocks that pass through to the price level) and thus allows the price level and money to wander but without affecting trend inflation.[22] The central bank can accommodate inflation shocks as long as it is credible. Specifically, the central bank can target core inflation (inflation stripped of volatile series like food and energy) while assuming that expected trend inflation remains unchanged. That is, the public does not extrapolate variability in observed inflation into the future. Subject to credibility, the central bank's reaction function causes nominal money demand to grow at a rate that does not require the inflation rate to differ from its target. All changes in money are price-preserving.

3. THE NONMONETARY VIEW OF INFLATION

The term "quantity theory" focuses on the kind of analytical framework useful for understanding the behavior of the price level by directing attention toward the way in which the central bank controls money creation. Trivially, as made

[21] The behavior of the economy is determined by its real business cycle core.

[22] Depending on the time-series properties of inflation shocks, inflation exhibits both persistence and variability around trend. It is important not to confuse that observed persistence (positive autocorrelation) in inflation with intrinsic (hard-wired) inflation. It does not follow that the central bank is reducing the variability of output by increasing the variability of inflation. At the same time, if the central bank attempted to eliminate transitory fluctuations in inflation around trend, it would increase the variability of output. Credibility allows it to control inflation without adding variability to output beyond what is built into the response of the real business cycle core of the economy to shocks.

evident by the discussion of the equation of exchange (1), real factors affect the price level. In contrast to the quantity-theory view, nonmonetary views make these real factors into the central actors determining the price level. In the form of an inflation shock, they raise the price level. A built-in rigidity in prices allows the central bank to reduce growth in real expenditure by lowering growth in nominal expenditure. As a result, it raises the unemployment rate. The central bank controls inflation by playing off one real factor (an increase in the unemployment rate) against another real factor (an inflation shock). According to this view, the central bank faces a menu of choices whereby it can reduce the variability of inflation by increasing the variability of unemployment, and conversely.

Here, I review the nonmonetary view that is associated with the traditional Keynesian Phillips curve (2). This variant shaped the policymaking environment in the stop-go period, which lasted from 1965 until 1979. The inflation rate is π_t. The output gap, x_t, is the difference between the log of actual output, y_t, and potential output, y_t^p, or $(y_t - y_t^p)$. To give the output gap empirical content, practitioners of this view often use as a proxy the cyclical behavior of output measured by the difference between actual output and a trend line fitted to output. The ε_t is an inflation or cost-push shock.

$$\pi_t = \pi_{t-1} + \alpha x_{t-1} + \varepsilon_t \qquad \alpha > 0 \tag{2}$$

From the perspective of the nonmonetary view, explanations of inflation are eclectic in the sense that each episode of inflation can possess its own primary cause. In the stop-go period, discussions of inflation typically began with a taxonomic classification of the different generic causes of inflation. The major classifications in this taxonomy were aggregate demand (demand-pull) and aggregate supply (cost-push), with propagation of these sources of inflation through intrinsic inflation persistence (a wage-price spiral).[23]

Demand-pull inflation arises from a positive output gap ($x_t > 0$). A variety of influences can boost real aggregate demand. At least through the early 1970s, the consensus among economists was that deficit spending (the full-employment surplus or deficit) exercised a strong influence on real aggregate demand while monetary policy actions, which worked through the interest rate, exercised only a negligible impact. Cost-push inflation arises from positive inflation shocks ($\varepsilon_t > 0$), that is, from factors that affect supply and demand in particular markets. Economists have identified inflation shocks with a large number of factors such as food and energy prices, depreciation of the foreign exchange value of the currency, monopoly power of unions and corporations, and government regulations. As reflected in the value of 1 on the coefficient on the π_{t-1} term, intrinsic inflation persistence propagates

[23] References are legion in the pre-1980 literature. See, for example, Ackley 1961 and Bronfenbrenner and Holzman 1963. See also Hetzel (2008, chaps. 1, 6, 11, 22, and 26).

these shocks unless the central bank offsets them by creating a negative output gap.[24] In the 1970s, economists often attributed inflation to a wage-price spiral set off by the aggregate-demand shock of Vietnam War spending and later the supply shocks of OPEC oil price increases (Nelson 2005 and Hetzel 2008, chaps. 6, 11, and 22).

The nonmonetary view has evolved over time. The dominant pre-1970s view did not associate the central bank with inflation. That changed after the association of inflation and high rates of money growth in the 1970s (Hetzel 2008, chap. 1). The prevailing view then changed to acceptance of the view that central banks can control inflation. However, the assumption was that to avoid a socially unacceptable high unemployment rate the central bank had to accommodate through high money growth the inflation caused by cost-push shocks. The genesis of inflation lies in excessive growth of real aggregate demand or in inflation shocks with hard-wired (intrinsic) propagation of the resulting inflation into future inflation, unless the central bank offsets it by raising unemployment. The central bank then faces a tradeoff. It can reduce inflation but only by increasing unemployment. More generally, the central bank can reduce the variability of inflation but only by increasing the variability of unemployment.

4. LEARNING FROM EXPERIENCE

Knowledge of what monetary policies the Fed followed in the past and of how they changed over time aids in the choice between the quantity theory and the nonmonetary view as the better description of how central banks control inflation. The reason is that each of these two views possesses different criteria for the success of monetary policies. According to the quantity-theory view, a monetary policy will work well only if it provides a nominal anchor and allows the price system to determine real variables. From the nonmonetary view, a successful monetary policy requires that policymakers choose an appropriate tradeoff between output (unemployment) variability and inflation variability, given the inflation shocks they confront. Also, policymakers need to achieve an optimal policy mix. Specifically, they should choose the optimal mix among monetary, fiscal, and incomes policies given their assessment of the nature of inflation as demand-pull, cost-push, or wage-spiral.[25]

Monetary policies have evolved with changes in the intellectual and political environment and also with the intellectual temper of FOMC chairmen

[24] In terms of the Phillips curve (2), the central bank would need to raise the real interest rate to reduce aggregate real demand, thereby creating a negative output gap ($x_t < 0$). A negative output gap would offset the positive effect of an inflation shock ($\varepsilon_t > 0$) on inflation (π_t).

[25] "Incomes policies" is the general term for government intervention in the price and wage setting of private markets.

(Hetzel 2008, chap. 2). Modern central banking began with the Treasury-Fed Accord of March 1951. In the changed intellectual environment of the post-war period, monetary policymakers replaced their assumed responsibility under the real bills doctrine to prevent what in their judgment constituted unsustainable increases in asset prices (due to speculation in stock and commodity markets) with responsibility for economic stabilization (Hetzel 2008, chaps. 3, 4, and 5). After the Accord, FOMC chairman William McChesney Martin created a monetary policy that adumbrated that of the Volcker-Greenspan era.[26]

Two major events shaped the monetary policy invented by Martin (and his advisor Winfield Riefler). First, with the 1953–1954 recession, the FOMC began to move the funds rate in a measured, persistent way in response to changes in the economy's rate of resource utilization. Second, when price stability ceded to inflation in the period from mid-1956 through 1958 and with the inflation scare of the summer of 1958, Martin began to move short-term interest rates promptly after cyclical turning points. In the spirit of real bills, his purpose was to prevent "speculation" in the financial markets. However, Martin made a momentous change. He directed monetary policy toward preventing the emergence of an inflation premium in bond markets rather than attempting to prevent what in policymakers' eyes constituted an unsustainable increase in asset prices (Hetzel 2008, chap. 5).[27] The Martin FOMC's reaction function, termed here LAW with credibility, foreshadowed that of Volcker-Greenspan (Hetzel 2008, chap. 21).

After the mid-1960s, monetary policy changed with the advent of stop-go.[28] With stop-go, the FOMC attempted to control the growth rate of real aggregate demand in a way that balanced the objectives of full employment and inflation. The appellation, stop-go, came from the practice of pursuing stimulative monetary policy during economic recoveries and restrictive policy later

---

[26] See Hetzel and Leach 2001a and 2001b; see also the link, "The Fiftieth Anniversary of the Treasury-Fed Accord" on http://www.richmondfed.org/publications/economic_research. The economics profession understood monetary policy in the context of aggregate-demand management with inflation arising as a consequence of the extent to which the level of aggregate demand stressed resource utilization. Not until the early 1970s did the economics profession assign a significant role to monetary policy as a determinant of aggregate real demand and, thus, as a useful tool for aggregate-demand management. In contrast, Martin understood the control of inflation in terms of the control of credit where the inflationary expectations of financial markets were a gauge of whether the extension of credit was excessive (Hetzel 2005, chap. 5).

[27] During the summer of 1958 and as seen later in 1983 and 1984, the FOMC looked for sharp, discrete increases in the bond rate as a proxy for an increase in expected inflation.

[28] Stop-go began in the Johnson administration. After the passage of the Kennedy tax cut in February 1964, both Congress and the administration united in their opposition to interest rate increases on the grounds that the increases would thwart the expansionary impact of the tax cuts. When inflation rose starting in 1965 and with his own house divided because of the appointment of governors by Democratic presidents Kennedy and Johnson, Martin opted for the use of monetary policy as a bargaining chip. If Congress would pass a tax surcharge, Martin would limit interest rate increases. Fiscal restraint, Martin hoped, would obviate the need for rate increases (Bremner 2004; Hetzel 2008, chap. 7).

on as inflation rose. How did stop-go alter the LAW-with-credibility procedures developed by the Martin FOMC (prior to the populist political pressures that arose during the Johnson administration)? The attempt during business cycle recoveries to lower unemployment (reduce the magnitude of the negative output gap) caused the FOMC to put inertia into short-term interest rates relative to cyclical movements in real output. The FOMC raised interest rates only belatedly after cyclical troughs when the unemployment rate was still high. Similarly, it lowered interest rates only slowly after cyclical peaks. As a result, money growth became pro-cyclical—rising and high during economic recovery and falling and low during recessions. With a lag, inflation followed these changes in money growth (Hetzel, chaps. 23–25). In go phases, the presumption was that a negative output gap (high unemployment) would allow monetary policy to be stimulative without raising inflation. In stop phases, the presumption was that a moderate negative output gap would allow a reduction in inflation at a socially acceptable cost in terms of unemployment—the policy of gradualism (Hetzel 2008, chaps. 7 and 8).

Stop-go arose from a conjunction of a political environment that demanded uninterrupted high real growth and low unemployment with an intellectual environment promising that government aggregate-demand policies could deliver these objectives. The Keynesian consensus held that the optimal combination of fiscal and monetary policy could deliver sustained real growth and high output while incomes policies could limit the resulting inflation (Samuelson and Solow [1960]1966). As manifested in beliefs about monetary policy, that consensus rested on two key premises. First, the price system does not work well to maintain full employment. From 1958 through 1965, excess unemployment (a negative output gap) apparently appeared in the form of an unemployment rate well above the assumed full-employment rate of 4 percent. Second, the price level is a nonmonetary phenomenon with inflation engendered at various times by either excess aggregate demand (demand-pull) or supply shocks (cost-push) and propagated by inflationary expectations untethered by monetary policy (a wage-price spiral).

This hard-wired (intrinsic) propagation of inflation supposedly imparted inertia to inflation relative to changes in aggregate nominal demand. Inertia in actual and expected inflation allows the central bank to exercise discretionary control over real variables (such as unemployment) through its control of aggregate nominal demand (expenditure). However, the downside of this inflation inertia is that the central bank has to create a significant amount of excess unemployment to offset the effects of inflation shocks and to maintain low, stable inflation. Because of this assumption, policymakers generally did not believe that monetary restriction was the socially optimal way of controlling inflation. Given the consensus that the inflation of the 1970s resulted from cost-push shocks propagated by a wage-price spiral, with the exception

of the Ford administration, all the presidential administrations from Kennedy through Carter used some form of incomes policies to control inflation.

Note the importance of the interaction between the above two premises about the inefficacity of the price system and the nonmonetary character of the price level. In a series of articles, Orphanides (for example, Orphanides 2002) documented the widespread belief during the 1970s that the unemployment rate exceeded its full-employment level (or the NAIRU, the non-accelerating inflation rate of unemployment). Using a Taylor rule framework, Orphanides (2003) attributed the inflation of the 1970s to this misestimation of the output gap. But why did policymakers not promptly revise their estimate of full employment with the first appearance of inflation? The reason is that they attributed inflation to cost-push factors. The assumed ability to parse the origin of inflation and decide whether an aggregate-demand policy or an incomes policy constituted the appropriate response was a far more fundamental failure than the technical issue of estimating the NAIRU correctly.

In the stop-go period, policymakers understood monetary policy as requiring the exercise of ongoing discretion about the socially acceptable level of unemployment to allow and, as a consequence, what amount of inflation to tolerate (Burns 1979; Hetzel 1998 and 2008, chap. 8). The presumed necessity of raising the unemployment rate to reduce an inflation rate assumed driven by cost-push shocks and propagated by a wage-price spiral appeared to demand discretion to manage adverse political reaction (Burns 1979). While a hard-wired inertia in inflation and inflationary expectations appeared to allow for this discretionary control of real variables, such inertia made the excess-unemployment cost of controlling inflation appear very high. Discretion, however, meant that nothing in central bank procedures imposed constancy of a nominal variable (such as stable long-run money growth) as a way of disciplining period-by-period funds rate changes to assure the time-consistency of policy (Hetzel 2008, chap. 1).[29]

The experiment with the discretionary juggling of unemployment and inflation targets caused expectations to change in a way that eventually vitiated the ability of the central bank to control real variables such as unemployment. The United States had entered into the period of stop-go policy from an environment of expected price stability created by the long experience with a commodity standard and, after the 1951 Treasury-Fed Accord, a monetary policy focused on price stability (Hetzel 2008, chaps. 4–7). For this reason, initially, the expansionary policy followed in the go phases of stop-go exerted a positive influence on real output. However, over business cycles, the FOMC allowed the inflation rate to drift upward (Hetzel 2008, chaps. 7,

[29] As the 1970s progressed, some regional Reserve Banks (San Francisco, Richmond, Philadelphia, and Minneapolis) joined St. Louis in arguing that the control of inflation required control of money growth.

8, 11, and 23–25). In 1966, when stimulative monetary policy began to raise inflation, the contemporaneous expectation that inflation was stationary (fluctuated around an unchanged base) allowed both inflation to increase without an increase in expected inflation and output to rise above trend. After 1967, this assumption of stationarity began to diminish until in 1979, it disappeared.[30] In 1979, the public began to associate inflation with the Fed rather than with the market power of large corporations and unions and with special factors affecting markets for energy, food, medical services, and so on (Hetzel 2008, chap. 12).[31] By 1979, inflationary expectations had neutralized the ability of monetary policy to stimulate the economy (Hetzel 2008, chaps. 1, 7, 8, 11, 13, 14, and 26).[32] Stop-go created the expectational environment described in Kydland-Prescott (1977) and Barro-Gordon (1983) in which the anticipatory behavior of price setters neutralizes the ability of monetary policy to control real output systematically.

To understand the completeness of the breakdown of the ability of policymakers to exploit Phillips curve tradeoffs, it is useful to recall statements by past policymakers. In perhaps the most famous statement summarizing the failure of aggregate-demand policies to control unemployment, James Callaghan, British Prime Minister, summarized the British experience in 1976 (cited in Nelson 2001, 27 and Wood 2005, 387):

> The cozy world we were told would go on forever, where full employment would be guaranteed by a stroke of the chancellor's pen, cutting taxes, deficit spending . . . is gone. . . .We used to think that you could spend your way out of a recession. . . .I tell you in all candour that that option no longer exists, and in so far as it ever did exist, it worked on each occasion since the war by injecting a bigger dose of inflation into the economy, followed by a higher level of unemployment as the next step.

---

[30] When inflation rose in 1966, initially monetary policy turned restrictive. However, unlike 1957 and 1958 when the Fed stayed with restriction until it had eliminated inflation, in 1967 it backed off (see fn. 28 and Hetzel 2008, chap. 7).

[31] The reason this recognition occurred only slowly was that the public faced the same sorts of problems faced by econometricians making inferences with a small number of observations. There were three sustained surges in inflation. The first followed the Vietnam War and inflation had always risen in war time. The second surge, which began in early 1973, could be explained by special factors dependent on supply shortages in oil, food, etc. The fact that trend inflation remained at about 6 percent after the second surge could be explained by an intrinsic inflationary momentum (the wage-price spiral). Only with the third surge that began in 1978 did any significant part of the economics profession or the public become receptive to Friedman's monetarist explanation for inflation that highlighted high rates of money creation.

[32] Lucas (1996, 679) wrote: "The main finding that emerged from the research in the 1970s is that. . . anticipated monetary expansions. . . are not associated with. . . stimulus to employment and production. . . .Unanticipated monetary expansions on the other hand can stimulate production as, symmetrically, unanticipated contractions can induce depression."

Volcker (12/3/80, 4) observed:

> [T]he idea of a sustainable "trade off" between inflation and prosperity . . . broke down as businessmen and individuals learned to anticipate inflation, and to act in this anticipation. . . .The result is that orthodox monetary or fiscal measures designed to stimulate could potentially be thwarted by the self-protective instincts of financial and other markets. Quite specifically, when financial markets jump to anticipate inflationary consequences, and workers and businesses act on the same assumption, there is room for grave doubt that the traditional measures of purely demand stimulus can succeed in their avowed purpose of enhancing real growth.

Greenspan (U.S. Cong. 2/19/93, 55–6) later made the same point:

> The effects of policy on the economy depend critically on how market participants react to actions taken by the Federal Reserve, as well as on expectations of our future actions. . . .[T]he huge losses suffered by bondholders during the 1970s and early 1980s sensitized them to the slightest sign . . . of rising inflation. . . .An overly expansionary monetary policy, or even its anticipation, is embedded fairly soon in higher inflationary expectations and nominal bond yields. Producers incorporate expected cost increases quickly into their own prices, and eventually any increase in output disappears as inflation rises.

The Volcker (12/3/80, 4) quotation above expresses the situation that he inherited upon becoming FOMC chairman in August 1979 (see also Goodfriend and King 2005; Lindsey, Orphanides, and Rasche 2005; and Hetzel 2008, chaps. 1, 13, and 26). Expected inflation had become positively related both to actual inflation and to above-trend real growth. Expected inflation passed through quickly to actual inflation. By 1979, the Fed was left with very little ability to produce a wedge between actual and expected inflation and, as a result, with very little ability to manipulate excess unemployment or an output gap.

Upon becoming FOMC chairman in August 1979, Volcker turned to money targets as a device for achieving credibility. Especially, Volcker hoped, the commitment to maintaining moderate money growth would convince the public that the FOMC would break the prior pattern of allowing inflation to rise during cyclical recoveries. However, the interest sensitivity of the demand for M1 (the monetary aggregate targeted by the FOMC) produced by the 1980 deregulation of deposit interest rates caused M1 velocity to become pro-cyclical (Hetzel and Mehra 1989). As a result, steady M1 growth would exacerbate cyclical fluctuations.

For this reason, in 1983 the FOMC moved to the LAW-with-credibility procedures originally foreshadowed by Martin. Measured by the inferred

behavior of the inflation premium in bond rates, the FOMC attempted to conduct policy in a way that produced low expected inflation consistent with low actual inflation. It also attempted to produce stable expected inflation in place of an expected inflation rate that rose in response to cyclically high real growth or inflation shocks. The effort by the Volcker-Greenspan FOMCs to reestablish the nominal expectational stability lost during the prior stop-go period finally succeeded in 1996. With the sharp increases in the funds rate in 1994 and early 1995, the Fed at last succeeded in allaying the fears of the bond market vigilantes, who had pushed up bond rates in response to above-trend real growth and inflation shocks (Hetzel 2008, chap. 15). Expected inflation ceased being a function of actual inflation and of above-trend real growth. For example, recently neither the recovery from the 2001 recession nor the sustained oil price shock that began in mid-2004 have raised expected inflation significantly above 2 percent as measured by the yield difference between nominal and TIPS (inflation-indexed) Treasury securities.

In the 1970s, a few economists (starting originally with Robert Lucas at Carnegie-Mellon and later at the University of Chicago) argued that the stagflation of the 1970s (the persistence of inflation despite assumed excess unemployment) resulted not from cost-push inflation but rather from the way that monetary policy conditioned inflationary expectations.[33] That is, it resulted from a lack of central bank credibility. Like the monetarists in the 1950s and 1960s, these economists constituted a miniscule minority of the profession. However, the success of the Volcker policy of disinflation changed dramatically the intellectual environment. Under Volcker, as a result of a focus on expected inflation, the FOMC simply accepted responsibility for inflation without regard to its presumed origin as aggregate-demand or cost-push (Hetzel 2008, chaps. 13 and 14). The desire to establish the credibility required to control expected inflation imposed overall consistency on monetary policy (Hetzel 2008, chap. 26). The demonstrated ability of monetary policy not only to control inflation but also to do so without periodic recourse to "high" unemployment gave credence to the idea that the central bank could control inflation through consistent application of policy thought of as a strategy. The application to monetary policy of the ideas of rational expectations by Lucas (1972, 1976, and 1980) and of rules by Kydland and Prescott (1977) went from being an intellectual curiosity to part of mainstream macroeconomics.

5. QUANTITY THEORY VERSUS THE NONMONETARY VIEW

Volcker and Greenspan resurrected Martin's policy of LAW with credibility in the form of "inflation targeting," in which the term does not refer to an

[33] Lucas (1972) developed the idea of rational expectations to undergird the idea that the central bank cannot systematically control real variables.

explicit inflation target but rather to policy procedures that keep trend inflation constant at a low level. Which view—the quantity-theory view or the nonmonetary view—provides the better framework for understanding the success of this policy? That is, how did the Volcker and then the Greenspan FOMCs discipline the "measured, persistent" changes in the funds rate made in response to sustained changes in the degree of resource utilization to maintain trend inflation unchanged in response to aggregate-demand shocks?

The quantity-theory view suggests an interpretation of the Volcker-Greenspan procedures in terms of what I call a "classical dichotomy." Credibility creates an expectational environment in which firms set prices consistent with unchanged trend inflation. Changes in the real funds rate then track the natural rate and allow the price system to determine real variables.

According to the nonmonetary view, the FOMC manipulates excess unemployment (an output gap) to manage inflation and inflation variability according to tradeoffs summarized by a Phillips curve. However, the experience with stop-go was not consistent with the existence of the required exploitable Phillips curve. The problem was that inflationary expectations changed in a way that offset the attempted control of real variables. It follows that if the central bank cannot manipulate the inflation rate to control unemployment then it also cannot manipulate unemployment to control inflation.

Moreover, the nonmonetary view does not accord with the policy procedures of the Volcker-Greenspan FOMCs. According to the nonmonetary view, periodic inflation shocks cause inflation to overshoot the central bank's (implicit) inflation target. There is a fixed sacrifice ratio, which is defined as the excess-unemployment cost of eliminating each percentage point of an inflation overshoot.[34] While the central bank can "stretch" the sacrifice ratio by eliminating inflation overshoots over long intervals of time, it must set a path for excess unemployment to constrain period-by-period funds rate changes such that the total of excess unemployment cumulates to the product of the inflation overshoot and the sacrifice ratio. However, nothing in the Volcker-Greenspan FOMC procedures corresponded to the treatment of excess unemployment as an intermediate target controlled as an intermediate step in controlling inflation (Hetzel 2008, chap. 21). Changes in the unemployment rate were merely an indicator of the change in the degree of resource utilization instead of an independent target.

[34] The number of man-years of unemployment in excess of full employment required to lower the inflation rate one percentage point.

6. CONCLUDING COMMENT

In the Volcker-Greenspan era, the desire of the Fed to reestablish the nominal expectational stability lost in the stop-go period produced rule-like behavior in the form of LAW with credibility. This policy separates the operation of the price system from the control of inflation—a classical dichotomy. Monetary policy relinquished determination of real variables to the price system while providing a stable nominal anchor in the form of low, stable expected inflation.

REFERENCES

Ackley, Gardner. 1961. *Macroeconomic Theory*. New York: The Macmillan Company.

Barro, Robert J., and David B. Gordon. 1983. "A Positive Theory of Monetary Policy in a Natural Rate Model." *Journal of Political Economy* 91 (August): 589–610.

Blinder, Alan S., and Ricardo Reis. 2005."Understanding the Greenspan Standard." Paper presented at the Federal Reserve Bank of Kansas City Economic Symposium, "The Greenspan Era: Lessons for the Future." Jackson Hole, WY.

Bremner, Robert P. 2004. *Chairman of the Fed: William McChesney Martin, Jr. and the Creation of the American Financial System*. New Haven, CT: Yale University Press.

Bronfenbrenner, Martin, and Franklyn D. Holzman. 1963. "Survey of Inflation Theory." *American Economic Review* 53 (4): 593–661.

Burns, Arthur F. 1979. *The Anguish of Central Banking*. Belgrade, Yugoslavia: Per Jacobsson Foundation.

Fisher, Irving. 1896. *Appreciation and Interest*. New York: The Macmillan Company.

Friedman, Milton. [1968] 1969. "The Role of Monetary Policy." In *The Optimum Quantity of Money and Other Essays,* ed. Milton Friedman. Chicago: Aldine Publishing Company.

Friedman, Milton. [1969] 1969. "The Optimum Quantity of Money." In *The Optimum Quantity of Money and Other Essays,* ed. Milton Friedman. Chicago: Aldine Publishing Company.

Goodfriend, Marvin. 1987. "Interest Rate Smoothing and Price Level Trend-Stationarity." *Journal of Monetary Economics* 19 (May): 335–48.

Goodfriend, Marvin. 1993. "Interest Rate Policy and the Inflation Scare Problem." Federal Reserve Bank of Richmond *Economic Quarterly* 79 (1): 1–24.

Goodfriend, Marvin. 2000. "Overcoming the Zero Bound on Interest Rate Policy." *Journal of Money, Credit and Banking* 32 (November, Part 2): 1007–35.

Goodfriend, Marvin. 2004. "Monetary Policy in the New Neoclassical Synthesis: A Primer." Federal Reserve Bank of Richmond *Economic Quarterly* 90 (3): 3–20.

Goodfriend, Marvin. 2007. "How the World Achieved Consensus on Monetary Policy." *Journal of Economic Perspectives* 21 (Fall): 47–68.

Goodfriend, Marvin, and Robert G. King. 1997. "The New Neoclassical Synthesis." *NBER Macroeconomics Annual 1997*, eds. Ben S. Bernanke and Julio J. Rotemberg. Cambridge, MA: The MIT Press.

Goodfriend, Marvin, and Robert G. King. 2005. "The Incredible Volcker Disinflation." *Journal of Monetary Economics* 52 (July): 981–1015.

Hetzel, Robert L. 1987. "Henry Thornton: Seminal Monetary Theorist and Father of the Modern Central Bank." Federal Reserve Bank of Richmond *Economic Review* 73 (4): 3–16.

Hetzel, Robert L. 1998. "Arthur Burns and Inflation." Federal Reserve Bank of Richmond *Economic Quarterly* 84 (1): 21–44.

Hetzel, Robert L. 1999. "Japanese Monetary Policy: A Quantity Theory Perspective." Federal Reserve Bank of Richmond *Economic Quarterly* 85 (1): 1–25.

Hetzel, Robert L. 2004. "How Do Central Banks Control Inflation?" Federal Reserve Bank of Richmond *Economic Quarterly* 90 (3): 47–63.

Hetzel, Robert L. 2005. "What Difference Would an Inflation Target Make?" Federal Reserve Bank of Richmond *Economic Quarterly* 91 (2): 45–72.

Hetzel, Robert L. 2006. "Making the Systematic Part of Monetary Policy Transparent." Federal Reserve Bank of Richmond *Economic Quarterly* 92 (3): 255–90.

Hetzel, Robert L. 2007."The Contributions of Milton Friedman to Economics." Federal Reserve Bank of Richmond *Economic Quarterly* 93 (1): 1–30.

Hetzel Robert L. 2008. *The Monetary Policy of the Federal Reserve: A History*. Cambridge: Cambridge University Press.

Hetzel, Robert L., and Ralph F. Leach. 2001a. "The Treasury-Fed Accord: A New Narrative Account." Federal Reserve Bank of Richmond *Economic Quarterly* 87 (1): 33–55.

Hetzel, Robert L., and Ralph F. Leach. 2001b. "After the Accord: Reminiscences on the Birth of the Modern Fed." Federal Reserve Bank of Richmond *Economic Quarterly* 87 (1): 57–64.

Hetzel, Robert L., and Yash Mehra. 1989. "The Behavior of Money Demand in the 1980s." *Journal of Money, Credit, and Banking* 21 (November): 455–63.

Hume, David. [1752] 1956. "Of Interest." In *Political Discourses*. In *Writings on Economics*, ed. Eugene Rotwein. Madison, WI: University of Wisconsin Press.

Humphrey, Thomas M. 1974. "The Quantity Theory of Money: Its Historical Evolution and Role in Policy Debates." Federal Reserve Bank of Richmond *Economic Review* 60 (3): 2–19.

Humphrey, Thomas M. 1983a. "The Early History of the Real/Nominal Interest Rate Relationship." Federal Reserve Bank of Richmond *Economic Review* 69 (3): 2–10.

Humphrey, Thomas M. 1983b. "Can the Central Bank Peg Real Interest Rates? A Survey of Classical and Neoclassical Opinion." Federal Reserve Bank of Richmond *Economic Review* 69 (5): 12–21.

Humphrey, Thomas M. 1990. "Fisherian and Wicksellian Price-Stabilization Models in the History of Monetary Thought." Federal Reserve Bank of Richmond *Economic Review* 76 (3): 3–19.

Humphrey, Thomas M., and Robert E. Keleher. 1982. *The Monetary Approach to the Balance of Payments, Exchange Rates, and World Inflation*. New York: Praeger Publishers.

Kydland, Finn E., and Edward C. Prescott. 1977. "Rules Rather than Discretion: The Inconsistency of Optimal Plans." *Journal of Political Economy* 85 (June): 473–91.

Lindsey, David E., Athanasios Orphanides, and Robert H. Rasche. 2005. "The Reform of October 1979: How It Happened and Why." In *Reflections on Monetary Policy 25 Years After October 1979*. Federal Reserve Bank of St. Louis *Review* 87 (March/April): 187–235.

Lucas, Robert E., Jr. 1981. "Expectations and the Neutrality of Money" [1972]; "Econometric Policy Evaluation: A Critique" [1976]; "Rules, Discretion, and the Role of the Economic Advisor" [1980]. In *Studies in Business-Cycle Theory,* ed. Robert E. Lucas, Jr. Cambridge, Mass.: The MIT Press.

Lucas, Robert E., Jr. 1996. "Nobel Lecture: Monetary Neutrality." *Journal of Political Economy* 104 (August): 661–82.

McCallum, Bennett T. 1981."Price Level Determinacy with an Interest Rate Policy Rule and Rational Expectations." *Journal of Monetary Economics* 8 (November): 319–29.

McCallum, Bennett T. 1986."Some Issues Concerning Interest Rate Pegging, Price Level Determinacy, and the Real Bills Doctrine." *Journal of Monetary Economics* 17 (January): 135–60.

Nelson, Edward. 2001. "What Does the UK's Monetary Policy and Inflation Experience Tell Us About the Transmission Mechanism?" Discussion Paper No. 3047. Centre for Economic Policy Research, November. Available at: http://www.cepr.org/pubs/dps/DP3047 (accessed April 18, 2008).

Nelson, Edward. 2005. "The Great Inflation of the Seventies: What Really Happened?" *Advances in Macroeconomics* 5 (1): Article 3. Available at: http://www.bepress.com/bejm/advances/vol5/iss1/art3 (accessed April 18, 2008).

Neumann, Manfred J. M. 1997. "Monetary Targeting in Germany." In *Towards More Effective Monetary Policy,* ed. Iwao Kuroda. New York: St. Martin's Press.

Orphanides, Athanasios. 2002. "Monetary Policy Rules and the Great Inflation." *American Economic Review* 92 (May): 115–20.

Orphanides, Athanasios. 2003. "The Quest for Prosperity Without Inflation." *Journal of Monetary Economics* 50 (April): 633–63.

Patinkin, Don. 1965. *Money, Interest, and Prices*. New York: Harper and Row.

Pigou, Arthur. C. 1917. "The Value of Money." *Quarterly Journal of Economics* 32 (November): 38–65.

Pigou, Arthur C. 1927. *Industrial Fluctuations*. London: Macmillan.

Rich, Georg. 1987. "Swiss and United States Monetary Policy: Has Monetarism Failed?" Federal Reserve Bank of Richmond *Economic Review* 73 (3): 3–15.

Samuelson, Paul, and Robert Solow. [1960] 1966. "Analytical Aspects of Anti-Inflation Policy." In *The Collected Scientific Papers of Paul A. Samuelson,* ed. Joseph Stiglitz. Cambridge, Mass.: The MIT Press: 2 (102): 1336–53.

Sargent, Thomas J., and Neil Wallace. 1975. "Rational Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule." *Journal of Political Economy* 83 (April): 241–54.

Velde, François. 2004. "Poor Hand or Poor Play? The Rise and Fall of Inflation in the U.S." Federal Reserve Bank of Chicago *Economic Perspectives* (Quarter 1): 34–51.

Wicksell, Knut. [1898] 1965. "Interest and Prices." Translated by R. F. Kahn. Reprinted. New York: A. M. Kelley.

Wood, John H. 2005. *A History of Central Banking in Great Britain and the United States*. New York: Cambridge University Press.

Economic Quarterly—Volume 94, Number 2—Spring 2008—Pages 173–196

Limits to Redistribution and Intertemporal Wedges: Implications of Pareto Optimality with Private Information

Borys Grochulski

Traditionally an object of interest in microeconomics, models with privately informed agents have recently been used to study numerous topics in macroeconomics.[1] Characterization of Pareto-optimal allocations is an essential step in these studies, because the structure of optimal institutions of macroeconomic interest depends on the structure of optimal allocations. In models with privately informed agents, however, characterization of optimal allocations is a complicated problem, relative to models in which all relevant information is publicly available, especially in dynamic settings with heterogenous agents, which are of particular interest in macroeconomics.

The objective of this article is to characterize Pareto-optimal allocations in a simple macroeconomic environment with private information and heterogenous agents. We focus on the impact of private information on the implications of Pareto optimality. To this end, we consider two economies that are identical in all respects other than the presence of private information. In each

The author would like to thank Huberto Ennis, Ilya Faibushevich, Thomas Lubik, and Ned Prescott for their helpful comments. The views expressed in this article are those of the author and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] These topics include business cycle fluctuations (e.g., Bernanke and Gertler 1989); optimal monetary policy (Athey, Atkeson, and Kehoe 2005); unemployment insurance (Atkeson and Lucas 1995, Hopenhayn and Nicollini 1997, Stiglitz and Yun 2005); capital income and estate taxation (Kocherlakota 2005, Albanesi and Sleet 2006, Farhi and Werning 2006); disability insurance and social welfare (Golosov and Tsyvinski 2006, Pavoni and Violante 2007); social security design (Stiglitz and Yun 2005, Grochulski and Kocherlakota 2007); financial intermediation (Green and Lin 2003); and asset pricing (Kocherlakota and Pistaferri 2008).

economy, we fully characterize the set of all Pareto-optimal allocations. By comparing the structure of the sets of optimal allocations obtained in these two cases, we isolate the effect private information has on the implications of Pareto optimality.

The economic environment we consider is, on the one hand, rich enough to have features of interest in a macroeconomic analysis, and, on the other hand, simple enough to admit elementary, closed-form characterization of Pareto-optimal allocations, both with and without private information. The model we use is a stylized, two-period version of the Lucas (1978) pure capital income economy that is extended, however, to incorporate a simple form of agent heterogeneity. We assume that the population is heterogenous in its preference for early versus late consumption. In particular, we assume that a known fraction of agents are impatient, i.e., have a strong preference for consumption in the first time period, relative to the rest of the population. In the economy with private information, individual impatience is not observable to anyone but the agent. A detailed description of the environment is provided in Section 1.

In our analysis, we exploit the connection between Pareto-optimal allocations and solutions to so-called social planning problems, in which a (stand-in) social planner maximizes a weighted average of the individual utility levels of the two types of agents. These planning problems are defined and solved for both the public information economy and the private information economy in Section 2. The solutions obtained constitute all Pareto-optimal allocations in the two economies.

In the third section, we compare the Pareto optima of the two economies along two dimensions. First, we examine their welfare properties by comparing the utility levels provided to agents in the cross-section of Pareto-optimal allocations. The range of individual utility levels supported by Pareto optima in the private information economy turns out to be much smaller than that of the public information economy. In this sense, private information imposes limits to redistribution that can be attained in this economic environment. Then, we compare the structures of optimal intertemporal distortions, which are often called intertemporal wedges, across the Pareto optima of the two economies. With public information, all Pareto-optimal allocations are free of intertemporal wedges. In the economy with private information, we find Pareto-optimal allocations characterized by a positive intertemporal wedge, and others characterized by a negative intertemporal wedge. We close Section 3 with a short discussion of the implications of wedges for the consistency of Pareto-optimal allocations with market equilibrium outcomes, which are studied in many macroeconomic applications. Section 4 draws a brief conclusion.

1. TWO MODEL ECONOMIES

We consider two parameterized model economies. The two economies have the same preferences and technology. They differ, however, with respect to the amount of public information.

The following features are common to both economies. Each economy is populated by a unit mass of agents who live for two periods, $t = 1, 2$. There is a single consumption good in each period, c_t, and agents' preferences over consumption pairs (c_1, c_2) are represented by the utility function

$$\theta u(c_1) + \beta u(c_2),$$

where β is a common-to-all discount factor, and θ is an agent-specific preference parameter. Agents are heterogenous in their relative preference for consumption at date 1. We assume a two-point support for the population distribution of the impatience parameter θ. Agents, therefore, can be of two types. A fraction μ_H of the agents are impatient with a strong preference for consuming in period 1. Denote by H the value of the parameter θ representing preferences of the impatient agents. A fraction $\mu_L = 1 - \mu_H$ are agents of the patient type. Their value of the impatience parameter θ, denoted by L, satisfies $L < H$.[2]

In the economies we consider, the production side is represented by a so-called Lucas tree. We assume that the economy is endowed with a fixed amount of productive capital stock—the tree.[3] Each period, the capital stock produces an amount Y of the consumption good—the fruit of the tree. The consumption good is perishable—it cannot be stored from period 1 to 2. The size of the capital stock, i.e., the tree, is fixed: the capital stock does not depreciate nor can it be accumulated.

In our discussion, we will focus attention on a particular set of values for the preference and technology parameters. This will allow for explicit analytical solutions to the optimal taxation problem studied in this article. In particular, we will take

$$u(\cdot) = \log(\cdot),\ \beta = \frac{1}{2},\ H = \frac{5}{2},\ L = \frac{1}{2},\ \mu_H = \mu_L = \frac{1}{2},\ Y = 1. \quad (1)$$

Roughly, the model period is thought of as being 25 years. The value of the discount factor β of $\frac{1}{2}$ corresponds to the annualized discount factor of about 0.973. The fractions of the two patience types are equal, and preferences are logarithmic. With $\frac{H}{L} = 5$, we consider a significant dispersion of the

[2] A formulation of preferences with the two types having different discount factors would be equivalent.

[3] In our study of optimal allocations, we abstract from private ownership of capital. Given that (a) capital is publicly observable and seizable, and (b) the society does not value individual utilities differently on the basis of individual wealth, this abstraction has no bearing on the problem we study. That is, the set of Pareto optimal allocations does not depend on who holds wealth in the economy.

impatience parameter in the population. The per-period product of the capital stock is normalized to one.

The two economies we consider differ with respect to the scope of public knowledge of each agent's individual impatience parameter. In the first economy we consider, each agent's preference type is public information, i.e., it is known to the agent and everyone else. In the second economy, each agent's individual impatience is known only to himself.

2. PARETO-EFFICIENT ALLOCATIONS

An allocation in this environment is a description of how the total output (i.e., the economy's capital income Y) is distributed among the agents each period. We consider only type-identical allocations, in which all agents of the same type receive the same treatment. An allocation, therefore, consists of four positive numbers, $c = (c_{1H}, c_{1L}, c_{2H}, c_{2L})$, where $c_{t\theta}$ denotes the amount of the consumption good in period t assigned to each agent of type θ.

In this section, we describe the efficient allocations. We use the standard notion of Pareto efficiency applied to type-identical allocations. We say that an allocation c is Pareto-dominated by an allocation $\hat{c}$ if all types of agents are at least as well off at $\hat{c}$ as they are at c and some are strictly better off. In our model, allocation c is Pareto-dominated by an allocation $\hat{c}$ if

$$\theta u(\hat{c}_{1\theta}) + \beta u(\hat{c}_{2\theta}) \geq \theta u(c_{1\theta}) + \beta u(c_{2\theta})$$

for both $\theta = H, L$, and if

$$\theta u(\hat{c}_{1\theta}) + \beta u(\hat{c}_{2\theta}) > \theta u(c_{1\theta}) + \beta u(c_{2\theta})$$

for at least one θ. An allocation c is Pareto-efficient in a given class of allocations if c belongs to this class and is not Pareto-dominated by any allocation $\hat{c}$ in this class of allocations.

Pareto Optima in the Public Types Economy

In our economy with public preference types, resource feasibility is the sole constraint on the class of allocations that can be attained. An allocation is resource-feasible if the total amount consumed each period does not exceed the total available output. That is, in our model, allocation c is resource-feasible (RF) if for $t = 1, 2$,

$$\sum_{\theta=H,L} \mu_\theta c_{t\theta} \leq Y. \tag{2}$$

In the public types economy, therefore, we are interested in allocations that are Pareto-efficient in the class of all RF allocations. We will refer to such allocations as First Best Pareto optima.

Characterizing the Set of All First Best Pareto Optima

In order to find all First Best Pareto-optimal allocations, it will be useful to consider a social planning problem defined as follows:

First Best Planning Problem *For each* $\gamma \in [0, +\infty]$*, find an allocation* $c = (c_{1H}, c_{1L}, c_{2H}, c_{2L})$ *that maximizes the value of the welfare objective*

$$\gamma\,[Hu(c_{1H}) + \beta u(c_{2H})] + Lu(c_{1L}) + \beta u(c_{2L}), \tag{3}$$

subject to resource feasibility constraints (2).[4]

In this problem, which we will refer to as the First Best planning problem, γ represents the relative weight that the social welfare criterion (3) puts on the agents of type H. The constraint set of the First Best planning problem is defined by the RF constraints (2). It is easy to check that this constraint set is compact (i.e., closed and bounded). This, and the fact that the objective (3) is continuous, implies that a solution to the First Best planning problem exists for every $\gamma \in [0, +\infty]$. Also, since the RF constraints are linear in consumption, the constraint set is convex. The objective (3) is strictly concave for each $\gamma \in (0, +\infty)$. Thus, the First Best planning problem has a unique solution for every $\gamma \in [0, +\infty]$.[5] Denote this unique solution by $c^*(\gamma)$.

The social planning problem is a useful tool for welfare analysis due to the following result: If the set of all feasible allocations is convex and the utility functions of all agent types are strictly increasing and strictly concave, then every solution $c^*(\gamma)$ to the social planning problem is a Pareto optimum, and every Pareto optimum is a solution to the social planning problem for some $\gamma \in [0, +\infty]$.[6]

Because of the concavity of u and the convexity of the set of RF allocations, this result applies in our economy with public types. Thus, we can exploit the connection between the set of Pareto optima and the set of solutions to the First Best social planning problem. We will solve the social planning problem for each $\gamma \in [0, +\infty]$. The solutions we obtain, $c^*(\gamma)$, will determine the set of First Best Pareto optima as we adjust the value of γ between zero and infinity.

Since the First Best planning problem is concave for each $\gamma \in [0, +\infty]$, the solution $c^*(\gamma)$ is given by the necessary and sufficient first-order conditions

---

[4] Alternatively, we could write the social objective as $\alpha\,[Hu(c_{1H}) + \beta u(c_{2H})] + (1 - \alpha)\,[Lu(c_{1L}) + \beta u(c_{2L})]$, with $\alpha \in [0, 1]$. Our formulation (3) is equivalent when $\gamma = \alpha/(1 - \alpha)$. Thus, $\gamma = +\infty$ corresponds to $\alpha = 1$, i.e., the social objective under $\gamma = +\infty$ is given by $Hu(c_{1H}) + \beta u(c_{2H})$.

[5] The optima for $\gamma = 0$ and $\gamma = \infty$, trivially, are unique as well, with optimal allocations assigning all consumption respectively to type L and type H.

[6] The argument for this is entirely standard. See, e.g., section 16E of Mas-Colell, Whinston, and Green (1995).

of this problem. Thus, we can find the solution $c^*(\gamma)$ by taking the first-order conditions and solving for c. Denoting by ρ_t the Lagrange multiplier of the RF constraint at date $t = 1, 2$, the first-order conditions with respect to consumption are as follows:

$$\begin{aligned} \gamma H u'(c_{1H}) &= \rho_1 \mu_H, && (4)\\ L u'(c_{1L}) &= \rho_1 \mu_L, && (5)\\ \gamma \beta u'(c_{2H}) &= \rho_2 \mu_H, \textit{ and} && (6)\\ \beta u'(c_{2L}) &= \rho_2 \mu_L. && (7) \end{aligned}$$

The multipliers ρ_t must be strictly positive at the solution because the objective (3) is strictly increasing in consumption, i.e., both RF constraints bind. For each $\gamma \in [0, +\infty]$, the optimum $c^*(\gamma)$ is the solution to the system of equations consisting of the first-order conditions (4)–(7) and the RF constraints (2).

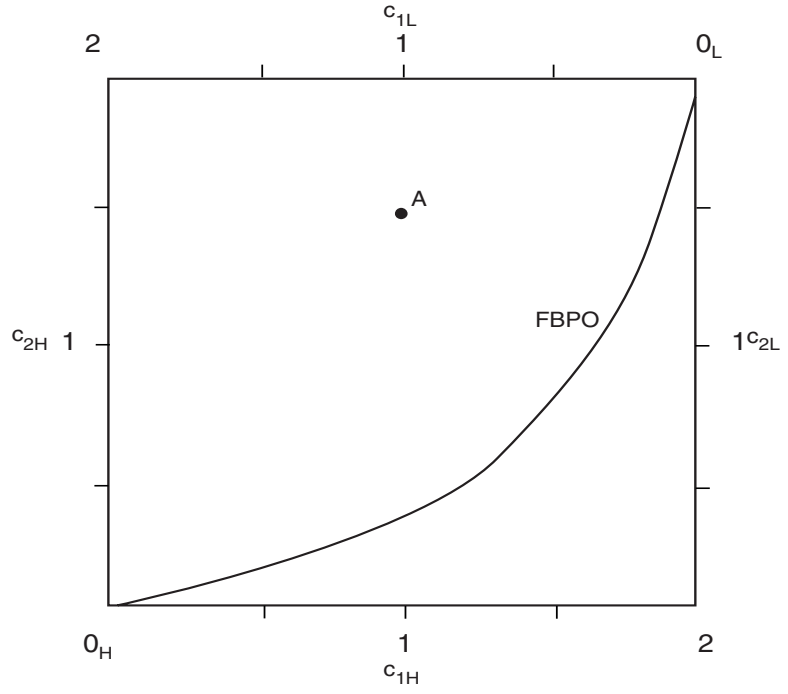
Using the parameterization (1), we can obtain a closed-form expression for the set of all First Best Pareto-optimal allocations, indexing the allocations in this set by γ. Solving for the optimal consumption values, as a function of γ, we get

$$\begin{aligned} c^*_{1H}(\gamma) &= \frac{10\gamma}{1+5\gamma}, && (8)\\ c^*_{1L}(\gamma) &= \frac{2}{1+5\gamma}, && (9)\\ c^*_{2H}(\gamma) &= \frac{2\gamma}{1+\gamma}, && (10)\\ c^*_{2L}(\gamma) &= \frac{2}{1+\gamma}. && (11) \end{aligned}$$

As we see, at any Pareto optimum, consumption allocated to the impatient type H is front-loaded, i.e., $c^*_{1H}(\gamma) > c^*_{2H}(\gamma)$, and consumption assigned to the less impatient type L is back-loaded, i.e., $c^*_{1L}(\gamma) < c^*_{2L}(\gamma)$. Looking across Pareto optima, consumption of the H-type is strictly increasing, at both dates, in the weight γ, while consumption of the L-type is strictly decreasing.

Figure 1 provides an Edgeworth-box representation of the set of all First Best Pareto optima. The Edgeworth box represents the set of all RF allocations at which the RF constraints (2) are satisfied as equalities (i.e., there is no waste of resources). In the Edgeworth box of Figure 1, the bottom-left corner represents the origin of measurement of consumption allocated to the agents of type H. The horizontal axis measures consumption in period 1. For example, point A in Figure 1, whose coordinates are (1, 1.5), represents an allocation at which the consumption of the H-type agents is $(c_{1H}, c_{2H}) = (1, 1.5)$.

Figure 1 The Set of First Best Pareto Optima



Note that since the fractions of the two types are equal and the resource constraints (2) are binding, we can write them as

$$c_{tL} = 2 - c_{tH} \tag{12}$$

for $t = 1, 2$. Thus, for a given consumption (c_{1H}, c_{2H}) allocated the H-type, the consumption allocated the L-type is given by

$$(c_{1L}, c_{2L}) = (2 - c_{1H}, 2 - c_{2H}).$$

Since the Edgeworth box represents only non-wasteful allocations, the top-right corner of the box of Figure 1, whose coordinates are (2, 2), is the origin of measurement of consumption allocated to the agents of type L. Point A in Figure 1, for example, represents an allocation that assigns amounts $(2 - 1, 2 - 1.5) = (1, 0.5)$ to the agents of type L.

The solid curve in Figure 1 represents the set of all First Best Pareto-optimal allocations given in (8)–(11). The allocations in this set are indexed by γ with the Pareto optimum for $\gamma = 0$ being in the bottom-left corner of the box, and the one obtained for $\gamma = \infty$ in the top-right corner. The curve representing the Pareto set is strictly increasing, which reflects the fact that

consumption of the H-type is strictly increasing in γ. As we noted before, for any weight γ, it is efficient to front-load consumption of the H-type and back-load consumption of the L-type. In the Edgeworth box of Figure 1, this is reflected by the fact that the First Best Pareto set lies below the 45 degree line (not depicted).

Pareto Optima in the Private Types Economy

In the second economy we consider, agents have private knowledge of their own impatience type θ. This imposes additional constraints on the set of allocations that are feasible in this environment.

As an example, suppose that the social planner—or simply the government—wants to distribute the total output of the Lucas tree according to the Pareto-optimal allocation $c^*(0) = (c^*_{1H}(0), c^*_{1L}(0), c^*_{2H}(0), c^*_{2L}(0)) = (0, 2, 0, 2)$. At this particular Pareto optimum, type H agents are assigned zero consumption in both periods (as the social welfare criterion (3) with $\gamma = 0$ does not value their utility at all), and agents of type L consume the whole output of the Lucas tree $Y = 1$. (Each agent of the L-type consumes 2 units, and the mass of the L-type agents is $\frac{1}{2}$, so the total consumption of the L-type agents is 1.) It is clear that when agents' types are private information, it is impossible for the government to attain this distribution of consumption. How will the government know which agent should be assigned zero consumption, as agents themselves are the only possible source of information about their preference type? If revealing the preference type H to the government means consuming zero in both periods, no impatient agent will admit being impatient. Thus, the Pareto optimum $c^*(0)$ is not feasible for the social planner when the impatience type is private information.

As this example demonstrates, the set of allocations feasible in the economy with private information is smaller than the set of all allocations satisfying the resource feasibility constraints (2). In particular, in addition to being resource-feasible, a feasible allocation of consumption c must also be incentive compatible. This requirement states that when faced with an allocation c, agents of both types must be willing to reveal truthfully their type to the government.[7]

Formally, an allocation $c = (c_{1H}, c_{1L}, c_{2H}, c_{2L})$ is incentive compatible (IC) if it satisfies the following two constraints:

$$Hu(c_{1H}) + \beta u(c_{2H}) \geq Hu(c_{1L}) + \beta u(c_{2L}) \tag{13}$$

and

$$Lu(c_{1L}) + \beta u(c_{2L}) \geq Lu(c_{1H}) + \beta u(c_{2H}). \tag{14}$$

[7] A general result known as the Revelation Principle (see Harris and Townsend 1981) guarantees that imposing the incentive compatibility requirement is actually without loss of generality.

Using this definition, we can simply say that the Pareto optimum $c^*(0)$ is not feasible in the economy with private types because it is not IC, as

$$\begin{aligned} Hu(c^*_{1H}(0)) + \beta u(c^*_{2H}(0)) &= Hu(0) + \beta u(0) \\ &< Hu(2) + \beta u(2) \\ &= Hu(c^*_{1L}(0)) + \beta u(c^*_{2L}(0)), \end{aligned}$$

and thus the IC constraint for the H-type, (13), is violated. The example of allocation $c^*(0)$ demonstrates that the set of feasible allocations in the private information economy is a strict subset of the set of allocations feasible in the public information economy. Moreover, this restriction on the feasibility is not irrelevant from the welfare perspective, as $c^*(0)$ is a Pareto optimum.

Characterizing the Set of Feasible Allocations with Private Types

Using the parameter values in (1), we can further characterize the set of feasible allocations in the private information economy, i.e., the set of all allocations that are RF and IC. Substituting the values in (1), the IC constraints (13) and (14) are given by, respectively,

$$\frac{5}{2}\log(c_{1H}) + \frac{1}{2}\log(c_{2H}) \geq \frac{5}{2}\log(c_{1L}) + \frac{1}{2}\log(c_{2L})$$

and

$$\frac{1}{2}\log(c_{1L}) + \frac{1}{2}\log(c_{2L}) \geq \frac{1}{2}\log(c_{1H}) + \frac{1}{2}\log(c_{2H}).$$

Using the RF constraints (12), we can eliminate from these inequalities consumption of the L-type agents. Simplifying and solving for c_{2H}, we obtain the following expressions for the IC conditions for the type H and L, respectively,

$$c_{2H} \geq \frac{2(2 - c_{1H})^5}{{c_{1H}}^5 + (2 - c_{1H})^5} \tag{15}$$

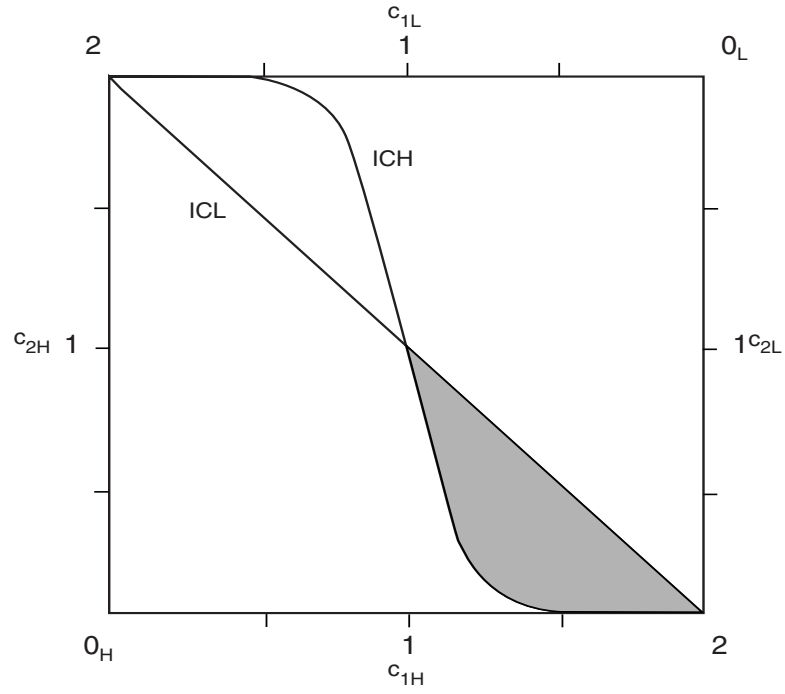
and

$$c_{2H} \leq 2 - c_{1H}. \tag{16}$$

Figure 2 depicts the set of all IC allocations in the Edgeworth box. The resource-feasible allocations that satisfy the IC constraint for type H, (15), lie on and above the curve ICH in Figure 2. Allocations that satisfy the IC constraint for type L, (16), lie on and below the line ICL. The shaded area, therefore, represents all IC allocations, i.e., those allocations that satisfy both IC conditions.

As we can see in Figure 2, the set of IC allocations (also satisfying the RF constraints as equalities) is convex. This property is not obvious a priori, as the IC constraints are given by nonlinear inequalities. Thus, the set of allocations

Figure 2 Incentive-Compatible Allocations in the Private Information Economy



feasible in the private information economy, i.e., those that satisfy the RF constraints as equalities and are incentive compatible, is convex.[8] Similar to the case of public information, this property is valuable as we can characterize the set of all Pareto optima in the private information economy by solving a planning problem.

Characterizing the Set of All Second Best Pareto Optima

Consider a planning problem defined as follows:

[8] Generally, the feasible set is not always convex in private information economies. Allocations involving lotteries over consumption bundles have been used in the literature to convexify the feasible set (see, e.g., Kehoe, Levine, and Prescott 2002). Also, when agents who misrepresent their type are more risk averse than those who report their type truthfully, lotteries may be welfare-improving even if the feasible set is convex (see Cole 1989). Neither of these reasons to consider lottery allocations, however, is present in the environment we consider in this article.

Second Best Planning Problem *For each $\gamma \in [0, +\infty]$, find an allocation $c = (c_{1H}, c_{1L}, c_{2H}, c_{2L})$ that maximizes the value of the welfare objective (3) subject to resource feasibility constraints (2) and incentive compatibility constraints (13), (14).*

Thanks to the convexity of the set of feasible allocations and the concavity of the objective, any solution to the Second Best planning problem is a Pareto optimum of the private information economy, and all such optima, referred to as the Second Best Pareto optima, can be obtained by solving this problem for all $\gamma \in [0, +\infty]$.[9]

Similar to the First Best planning problem, the Second Best planning problem is a concave maximization problem. Thus, for each γ, a unique solution exists. Let us denote this solution by $c^{**}(\gamma)$. As before, we can find $c^{**}(\gamma)$ using the first-order conditions.

There is, however, one difficulty in the private information economy that does not appear in the public information case: we do not know which, if any, IC constraints (13), (14) bind in the Second Best planning problem for a particular value of γ.

To determine which IC constraints bind for different values of γ, it will be helpful to return to the Edgeworth box. Figure 3 combines the curve representing the set of First Best Pareto optima from Figure 1, denoted by FBPO, with the set of IC allocations from Figure 2.

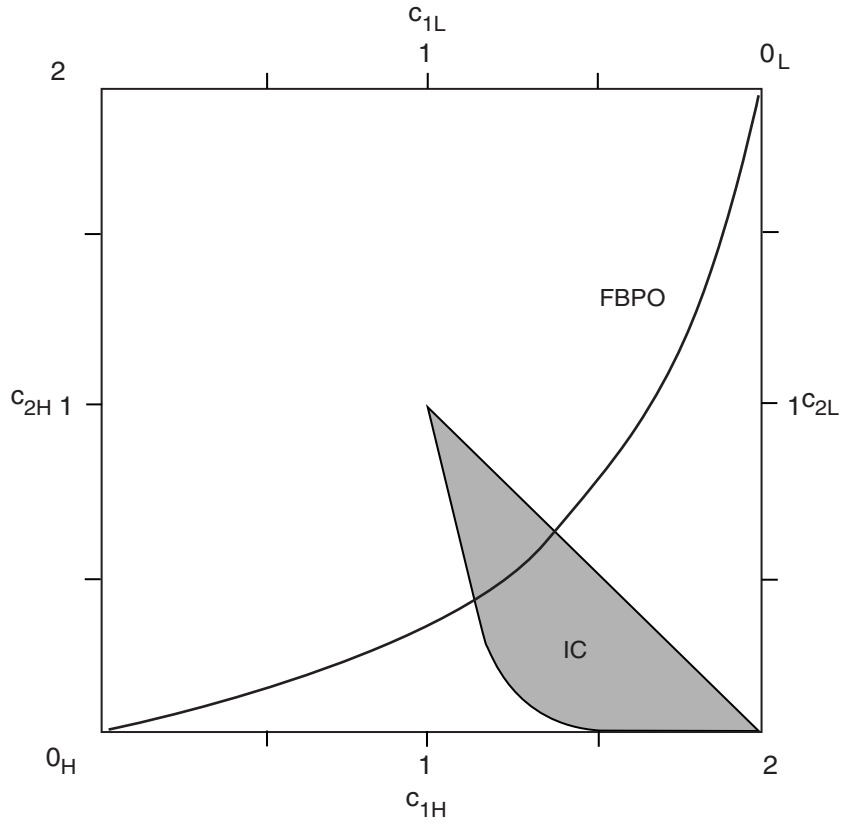
The first observation we make in Figure 3 is that a whole segment of the FBPO curve lies inside the IC set of the Second Best planning problem. Thus, for the values of the weight parameter γ for which the First Best Pareto optimum $c^*(\gamma)$ satisfies the IC constraints, the First Best Pareto optimum is also a solution to the Second Best planning problem, so $c^{**}(\gamma) = c^*(\gamma)$.

Second, we see that the First Best Pareto optima in the segment of the set FBPO that lies above the IC set are not incentive compatible because they violate the IC constraint of the L-type, (14). Similarly, the First Best Pareto optima in the segment of the set FBPO that lies below the IC set are not incentive compatible because they violate the IC constraint of the H-type, (13).

These observations suggest what the following lemma demonstrates formally. See the Appendix for a formal proof.

[9] Second Best Pareto optima are often referred to in the literature as constrained Pareto optima.

Figure 3 Incentive-Compatible Allocations and the First Best Pareto Optima



Lemma 1 *In the Second Best planning problem, we have the following. For all* $\gamma \in [\gamma_1, \gamma_2]$*, where*

$$\begin{aligned} \gamma_1 &= 5^{-\frac{5}{6}} \approx 0.26, \\ \gamma_2 &= 5^{-\frac{1}{2}} \approx 0.45, \end{aligned}$$

no IC constraints bind.
For all $\gamma > \gamma_2$*, the IC constraint of the L-type, (14), binds, and the IC constraint of the H-type, (13), does not.*
For all $\gamma < \gamma_1$*, the IC constraint of the H-type, (13), binds, and the IC constraint of the L-type, (14), does not.*

By Lemma 1, the Second Best Pareto optimum $c^{**}(\gamma)$ coincides with the First Best Pareto optimum $c^{*}(\gamma)$ for all welfare weights $\gamma \in [\gamma_1, \gamma_2]$. Also,

for $\gamma < \gamma_1$, the Second Best Pareto optimum $c^{**}(\gamma)$ can be found by solving a relaxed Second Best planning problem in which the IC of the L-type, (14), is dropped and the IC constraint of the H-type, (13), holds as equality. Similarly, for $\gamma > \gamma_2$, the Second Best Pareto optimum $c^{**}(\gamma)$ can be found by solving a relaxed Second Best planning problem in which the IC constraint of the H-type is dropped and the IC constraint of the L- type holds as equality.

Taking the first-order conditions of the relaxed Second Best planning problem for $\gamma > \gamma_2$, we obtain

$$(\gamma - \lambda_L \frac{L}{H}) H u'(c_{1H}) = \rho_1 \mu_H, \tag{17}$$

$$(1 + \lambda_L) L u'(c_{1L}) = \rho_1 \mu_L, \tag{18}$$

$$(\gamma - \lambda_L) \beta u'(c_{2H}) = \rho_2 \mu_H, \tag{19}$$

$$(1 + \lambda_L) \beta u'(c_{2L}) = \rho_2 \mu_L, \tag{20}$$

where $\lambda_L > 0$ is the multiplier on the IC constraint (14). For each $\gamma > \gamma_2$, the Second Best Pareto optimum $c^{**}(\gamma)$ is the solution to the system of equations consisting of the first-order conditions (17)–(20), the resource constraints (2), and the binding IC constraint (14). Using the parameter values in (1), we can solve explicitly for the optimum. After some algebra, we obtain

$$c^{**}_{1H}(\gamma) = c^{**}_{2L}(\gamma) = \frac{1 + 5\gamma}{1 + 3\gamma}, \tag{21}$$

$$c^{**}_{2H}(\gamma) = c^{**}_{1L}(\gamma) = \frac{1 + \gamma}{1 + 3\gamma}, \tag{22}$$

for all $\gamma > \gamma_2$. Similarly, taking the first-order conditions of the relaxed Second Best planning problem for $\gamma < \gamma_1$, we have

$$(\gamma + \lambda_H) H u'(c_{1H}) = \rho_1 \mu_H, \tag{23}$$

$$(1 - \lambda_H \frac{H}{L}) L u'(c_{1L}) = \rho_1 \mu_L, \tag{24}$$

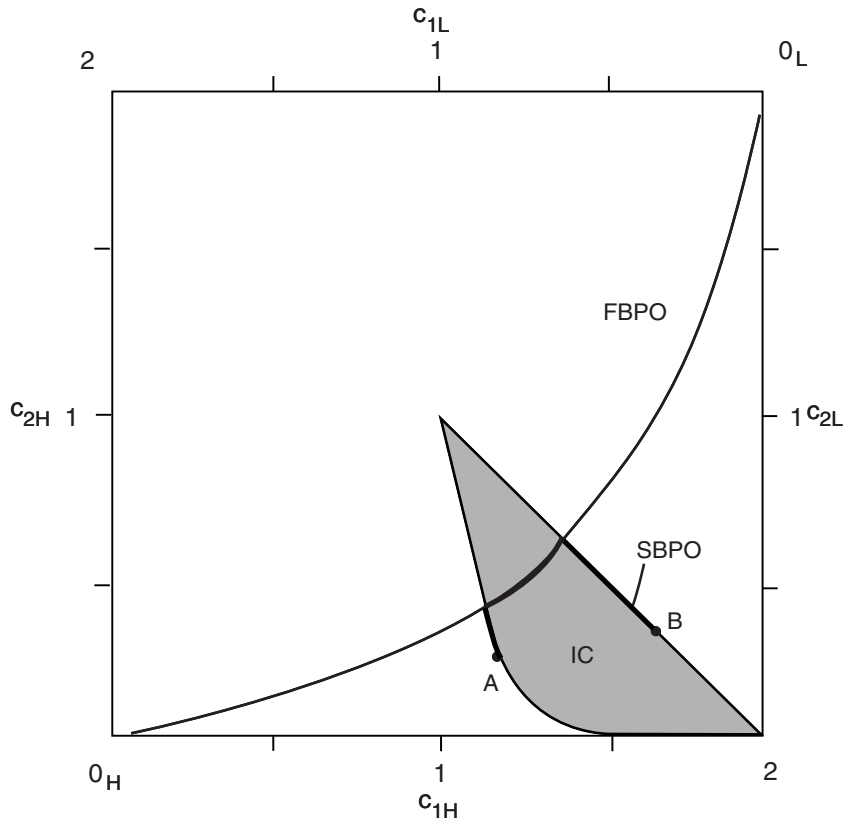
$$(\gamma + \lambda_H) \beta u'(c_{2H}) = \rho_2 \mu_H, \text{ and} \tag{25}$$

$$(1 - \lambda_H) \beta u'(c_{2L}) = \rho_2 \mu_L, \tag{26}$$

where $\lambda_H > 0$ is the multiplier on the IC constraint (13). Using the parameter values in (1), for each $\gamma < \gamma_1$, we can solve these first-order conditions, together with the resource constraints and the binding IC constraint, and obtain the Pareto optimum $c^{**}(\gamma)$.

Figure 4 represents the full set of Second Best Pareto-optimal allocations in the Edgeworth box. This figure also depicts the set of IC allocation and the set of First Best Pareto optima. For $\gamma \in [\gamma_1, \gamma_2]$, the Second and First Best Pareto optima coincide. The Second Best optima $c^{**}(\gamma)$ for $\gamma < \gamma_1$ lie on the lower edge of the IC set, where the IC constraint for the H-type binds. Point A represents the Second Best Pareto optimum $c^{**}(0)$. Similarly, the Second Best optima $c^{**}(\gamma)$ for $\gamma > \gamma_2$ lie on the upper edge of the IC set, where the

Figure 4 The Set of Second Best Pareto Optima



IC constraint for the L-type binds. Point B represents the Second Best Pareto optimum $c^{**}(\infty)$.

3. COMPARING PARETO OPTIMA IN THE TWO ECONOMIES

Having characterized the sets of optimal allocations in the public and private information economies, we can now compare their structures. In the first subsection, we compare the welfare properties of the two sets of Pareto optima. In the second subsection, we compare the structure of intertemporal wedges characterizing optimal allocations in the two economies.

Limits to Redistribution Under Private Information

Using the closed-form solutions we have obtained for the sets of First and Second Best Pareto optima, we can compute the value of utility optimally delivered to the two types of agents in the two economies. Denote by $V_\theta^*(\gamma)$ the lifetime utility delivered to each agent of type θ at the First Best Pareto optimum $c^*(\gamma)$ for $\gamma \in [0, \infty]$.[10] By $V_\theta^{**}(\gamma)$ denote the lifetime utility delivered to each agent of type θ at the Second Best Pareto optimum $c^{**}(\gamma)$ for $\gamma \in [0, \infty]$.

Figure 5 depicts the so-called First Best Pareto frontier. The concave curve represents the pairs of values $(V_H^*(\gamma), V_L^*(\gamma))$ for γ between 0.025 and 40. Outside this range, the frontier extends toward negative infinity and converges to a horizontal and vertical line. Point A in Figure 5 represents the values $(V_H^*(\frac{1}{3}), V_L^*(\frac{1}{3}))$. Point B marks the values $(V_H^*(1), V_L^*(1))$.

Figure 6 graphs the whole Second Best Pareto frontier, as well as a small section of the First Best frontier. The Second Best Pareto frontier consists of all points $(V_H^{**}(\gamma), V_L^{**}(\gamma))$ for $\gamma \in [0, \infty]$. As in Figure 5, points A and B represent the values $(V_H^*(\frac{1}{3}), V_L^*(\frac{1}{3}))$ and $(V_H^*(1), V_L^*(1))$. Because $1/3 \in [\gamma_1, \gamma_2]$, where First and Second Best Pareto optima coincide, point A belongs to the Second Best Pareto frontier. However, B lies outside of this set. The values $(V_H^{**}(1), V_L^{**}(1))$ are represented by point C in Figure 6.

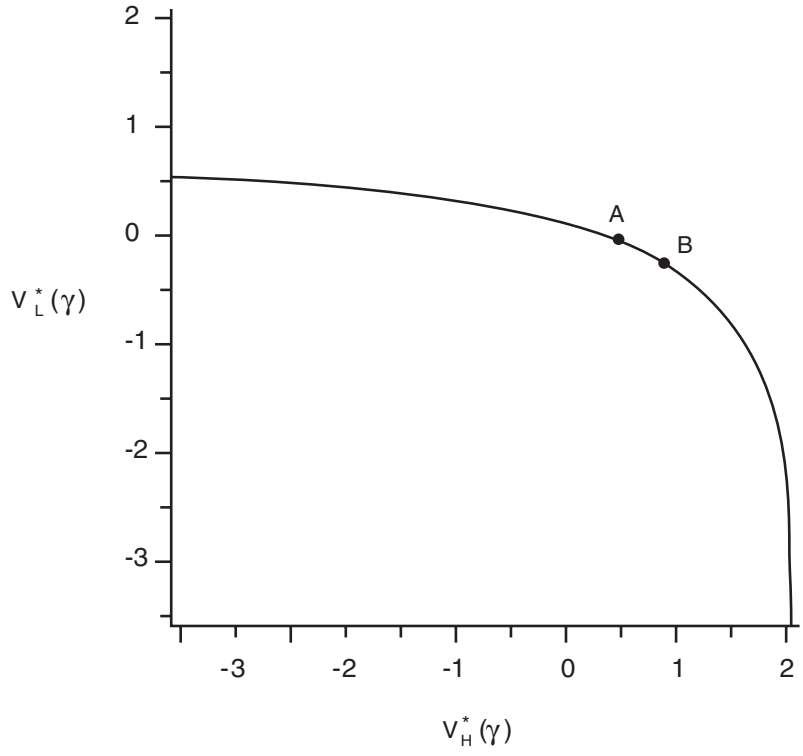
Comparing Figures 5 and 6, we note that private information severely restricts the range of the utility levels that can be provided to the two agent types, relative to the public information economy. With public information, the impatient type H can be provided with welfare as high as $V_H^*(\infty) = 2.08$, while under private information, the maximum welfare for the impatient type is $V_H^{**}(\infty) = 0.78$. For the agents of the patient type L, these maximal values are $V_L^*(0) = 0.69$ and $V_L^{**}(0) = 0.17$, respectively. Private information, thus, puts limits on the amount of redistribution that can be attained by a social planner.[11]

To gain some intuition on how these limits arise, we return to Figure 4 and consider the optimal allocation at the upper end of the range of γ for which private and public information optima coincide, i.e., $\gamma = \gamma_2$. The impact of private information on welfare attained in the two economies can be seen as we consider the values of $\gamma > \gamma_2$. In the public information economy, in order to increase welfare of the type H agents, the social planner simply increases consumption allocated to type H at both dates. That is, both $c_{1H}^*(\gamma)$ and $c_{2H}^*(\gamma)$ increase in γ, which of course means that both $c_{1L}^*(\gamma)$ and $c_{2L}^*(\gamma)$ decrease in γ. As γ grows, the consumption of the L-type becomes smaller

[10] That is, $V_\theta^*(\gamma) = \theta u(c_{1\theta}^*) + \beta u(c_{2\theta}^*)$ for $\theta = H, L$ and $\gamma \in [0, \infty]$.

[11] Redistribution is measured here in terms of utility, relative to a benchmark level, which does not have to be explicitly specified as the statement is true for any benchmark.

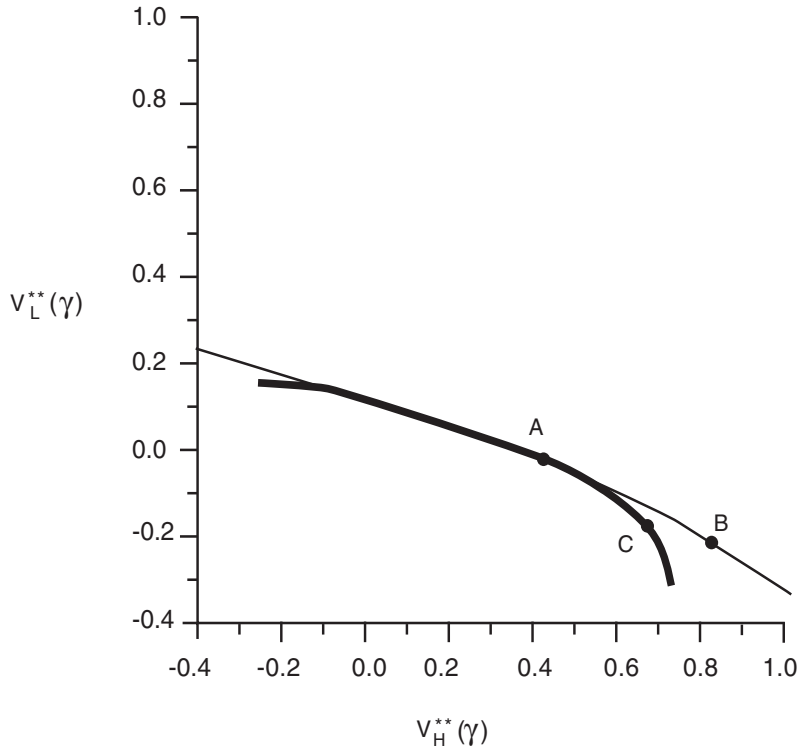
Figure 5 The First Best Pareto Frontier



and smaller. Resource feasibility is the only limit to this process. Eventually, the H-type consumes the economy's whole output.

Private information, however, puts a much more stringent limit on how small consumption of the agents of type L can be. At the Second Best optimum with $\gamma = \gamma_2$, consumption of the L-type is already small enough that the agents of type L are indifferent between their allocation and that intended for the H-type. Maximizing the social welfare criterion with $\gamma > \gamma_2$, the planner cannot improve the H-types' welfare by increasing its consumption at both dates, as this would violate the incentive compatibility condition for the L-type, i.e., the agents of type L would misrepresent their type. As γ is raised above γ_2, the planner increases H-types' welfare by increasing their consumption at date 1 and preserves incentive compatibility for the L-types by increasing their consumption at date 2. Because type H has a strong preference for consumption at date 1, relative to type L, it is possible to simultaneously compensate the L-type and increase the welfare of the H-type, to a point. In Figure 4, the Second Best Pareto optima $c^{**}(\gamma)$ for $\gamma > \gamma_2$

Figure 6 The First and Second Best Pareto Frontier



lie on the negative 45 degree line given by the upper edge of the set of IC allocations. At point B, which represents the Second Best optimum $c^{**}(\infty)$, the planner wants to further maximize H-types' welfare, regardless of type L's welfare. However, no further increase in H-types' welfare is possible. At $c_H^{**}(\infty) = (\frac{5}{3}, \frac{1}{3})$, the marginal utility levels of H-types' consumption at dates 1 and 2 are equal.[12] Adding one unit of consumption at date 1 and subtracting one unit of consumption at date 2 is not going to improve H-types' welfare. But in order to preserve incentive compatibility, the planner has to compensate any increase in H-types' consumption at date 1 with a one-to-one increase of L-types' consumption at date 2. Preserving incentive compatibility for the L-type, therefore, becomes too expensive for the planner to be able to further increase H-types' welfare. Thus, even though the social welfare objective does

[12] It is easy to check that $Hu'(c_{1H}^{**}(\infty)) = \beta u'(c_{2H}^{**}(\infty)) = \frac{3}{2}$.

not value the utility of L-type at all, it is not feasible in the private information economy to further redistribute to the H-type agents.

The same intuition applies to the limit that private information puts on the value that can be delivered to the L-type. As γ decreases below γ_1, the planner increases the utility of the L-types by increasing their consumption at date 2 and compensates the H-types with an increase in their consumption at date 1. At point A in Figure 4, the compensation for the H-type needed to preserve incentive compatibility becomes too large (and L-types' marginal utility of consumption at date 2 relative to marginal utility of consumption at date 1 too small) for a further increase in L-types' welfare to be feasible.

In Figure 6, we see that the presence of private information affects the value delivered to the disfavored type much more strongly than it affects the value delivered to the favored type, under any γ outside of $[\gamma_1, \gamma_2]$. When $\gamma = \infty$, the L-type consumes zero at the First Best Pareto optimum $c^*(\infty)$, i.e., $V_L^*(\infty) = -\infty$. In the private information economy, however, the L-type receives consumption $(\frac{1}{3}, \frac{5}{3})$ at the Second Best Pareto optimum $c^{**}(\infty)$, and $V_L^{**}(\infty) = -0.29 > -\infty$. Similarly, with $\gamma = 0$, welfare of the type H is $-\infty$ in the public information economy, but it is a finite number in the economy with private information.[13]

In addition, comparing points B and C in Figure 6, we see that when the social welfare objective is purely utilitarian, i.e., $\gamma = 1$, the L-type agents are better off in the private information economy. This observation generalizes. It is not hard to show that for all $\gamma > \gamma_2$, the disfavored L-types are better off when their type is private information, as in this case where the social planner's ability to redistribute to the H-type is hampered. Similarly, if $\gamma < \gamma_1$, i.e., when the H-types' utility receives a low weight in the social welfare criterion, we have that $V_H^{**}(\gamma) > V_H^*(\gamma)$, i.e., the disfavored H-type is better off in the private information economy.

Optimal Intertemporal Wedges

In order to gain further insight into the structure of the optimal allocations in the public and private information economies, we examine the intertemporal wedges in this subsection. Intertemporal wedge is defined as the difference between the social and the individual shadow interest rate. Wedges associated with a given Pareto optimum give us an understanding of the implicit distortions that are optimally imposed on the agents.

[13] The value of negative infinity is specific to the logarithmic utility. Under a constant relative risk aversion utility function with relative risk aversion smaller than one, for example, this value would be zero, i.e., a finite number. That the value delivered to the disfavored type is strongly impacted by the presence of private information remains true, however, for any strictly concave utility function.

We clarify the definitions as follows: the social shadow interest rate R^* associated with a Pareto-optimal allocation c^* is the number R at which the planner would choose to not alter the allocation c^* if given a chance to re-solve the social planning problem with access to a borrowing and savings technology with gross interest rate R. Similarly, the private shadow interest rate R^*_θ for $\theta = H, L$ is the number R at which the agents of type θ would not find it beneficial to trade away from their individual consumption allocation c^*_θ if they could borrow and save at the gross interest rate R.

In the simple economic environment that we consider, characterization of social and private shadow interest rates is straightforward. The social shadow interest rate is given by the ratio $\frac{\rho_1}{\rho_2}$ of the Lagrange multipliers associated with the resource feasibility constraints (2) at dates 1 and 2.[14] The private shadow interest rate of type θ at an optimum c^* is given by the ratio of type θ's marginal utility at date 1 and 2, i.e., $\theta u'(c^*_{1\theta})/\beta u'(c^*_{2\theta})$.[15]

Public Information Economy

Directly from the first-order conditions (4)–(7), we obtain that the First Best optima $c^*(\gamma)$ satisfy

$$\frac{\theta u'(c^*_{1\theta}(\gamma))}{\beta u'(c^*_{2\theta}(\gamma))} = \frac{\rho_1}{\rho_2},$$

for both $\theta = H, L$ and any $\gamma \in [0, \infty]$. The intertemporal wedge, given by the difference between the social and private shadow interest rate, is zero. This means that it is never optimal to distort the private intertemporal margin in the public information economy.

Private Information Economy

In the private information economy, the intertemporal wedges are zero at the Second Best Pareto optimum $c^{**}(\gamma)$ for any $\gamma \in [\gamma_1, \gamma_2]$, because the Second Best Pareto optimum $c^{**}(\gamma)$ coincides with the First Best Pareto optimum $c^*(\gamma)$ for each γ in this range.

---

[14] If the planner could borrow and lend at the gross interest rate R, the resource feasibility constraints of the social planning problem would be given by $\sum_\theta \mu_\theta c_{1\theta} + S \leq Y$ and $\sum_\theta \mu_\theta c_{2\theta} \leq Y + RS$, where S is the planner's saving at date 1. The first-order condition of this problem with respect to S is $-\rho_1 + R\rho_2 = 0$. This means that if $R = \rho_1/\rho_2$, the presence of the intertemporal saving technology does not alter the solution to the social planning problem, i.e., ρ_1/ρ_2 is the social shadow interest rate.

[15] This follows from the first-order condition with respect to individual savings s, evaluated at $s = 0$, of the individual optimal re-trading problem $\max_s \theta u(c^*_{1\theta} - s) + \beta u(c^*_{2\theta} + Rs)$.

For $\gamma > \gamma_2$, the first-order conditions in the Second Best planning problem, (17)–(20), imply that

$$\frac{Lu'(c_{1L}^{**}(\gamma))}{\beta u'(c_{2L}^{**}(\gamma))} = \frac{\rho_1 \mu_L/(1+\lambda_L)}{\rho_2 \mu_L/(1+\lambda_L)} = \frac{\rho_1}{\rho_2},$$

which means that an intertemporal wedge of zero is optimal for the agents of type L. From the same first-order conditions we obtain that

$$\frac{Hu'(c_{1H}^{**}(\gamma))}{\beta u'(c_{2H}^{**}(\gamma))} = \frac{\rho_1 \mu_H/(\gamma - \lambda_L \frac{L}{H})}{\rho_2 \mu_H/(\gamma - \lambda_L)} < \frac{\rho_1 \mu_H/(\gamma - \lambda_L)}{\rho_2 \mu_H/(\gamma - \lambda_L)} = \frac{\rho_1}{\rho_2},$$

which means that a strictly positive intertemporal wedge is optimal for the agents of type H at each Second Best Pareto optimum $c^{**}(\gamma)$ with $\gamma > \gamma_2$. The positive wedge means that agents of type H are *savings-constrained* at the optimal allocation of the private information economy when $\gamma > \gamma_2$. If agents could trade away from the optimum by borrowing or saving at the social shadow interest rate, the agents of type H would like to save. Note that the L-type agents would choose to not trade away from their consumption allocation, as their intertemporal wedge is zero.

The literature studying Pareto-optimal allocations in multi-period economies with private information finds that the positive intertemporal wedge characterizes Pareto-optimal allocations in many such environments.[16]

For $\gamma < \gamma_1$, the first-order conditions (23)–(26) of the Second Best planning problem imply that

$$\frac{Hu'(c_{1H}^{**}(\gamma))}{\beta u'(c_{2H}^{**}(\gamma))} = \frac{\rho_1 \mu_H/(\gamma + \lambda_H)}{\rho_2 \mu_H/(\gamma + \lambda_H)} = \frac{\rho_1}{\rho_2},$$

and

$$\frac{Lu'(c_{1L}^{**}(\gamma))}{\beta u'(c_{2L}^{**}(\gamma))} = \frac{\rho_1 \mu_L/(1 - \lambda_H \frac{H}{L})}{\rho_2 \mu_L/(1-\lambda_H)} > \frac{\rho_1 \mu_L/(1-\lambda_H)}{\rho_2 \mu_L/(1-\lambda_H)} = \frac{\rho_1}{\rho_2}.$$

This means that the optimal intertemporal wedge is zero for the H-type, and strictly negative for the L-type at any Second Best Pareto optimum $c^{**}(\gamma)$ with $\gamma < \gamma_1$. Therefore, we have that agents of type L are *borrowing-constrained* at the optimal allocation of the private information economy when $\gamma < \gamma_1$. If agents could borrow and lend at the social shadow interest rate, the L-type agents would like to borrow. This property is different from the intertemporal wedge typically found in the literature, in which, as we mentioned before, the positive intertemporal wedge is prevalent.

The intertemporal wedges associated with an optimal allocation give us an understanding of what distortions are optimal in agents' intertemporal consumption patterns. These distortions are relevant for the analysis of the welfare

[16] Articles that find this property of the optimal allocations include Diamond and Mirrlees (1978); Rogerson (1985); and Golosov, Kocherlakota, and Tsyvinski (2003).

properties of equilibrium outcomes in market economies. In a market economy, by definition, agents can use markets to trade away from the socially optimal allocation. Therefore, the negative intertemporal wedge in the optimal allocation for the L-type, which we have at any Pareto optimum $c^{**}(\gamma)$ with $\gamma < \gamma_1$, can be consistent with market equilibrium only if agents of type L can be prevented from borrowing at the social shadow interest rate. At the same time, however, any such disincentive to borrow cannot affect the agents of type H, whose private shadow interest rate is aligned with the social shadow interest rate at any optimum $c^{**}(\gamma)$ with $\gamma < \gamma_1$.

Detailed analysis of the issue of consistency between Pareto optima and market equilibria is beyond the scope of this article. This issue, however, plays an important role in the macroeconomic applications of private information models. It is central, for example, in the study of information-constrained optimal taxation problems.[17]

4. CONCLUSION

Our analysis of a simple macroeconomic environment with heterogenous agents provides an elementary exposition of the implications of Pareto optimality with private information. We obtain closed-form representation of all Pareto-optimal allocations with and without private information. We highlight the limits that private information puts on the utility distributions that can be attained in our environment. In addition, we provide a complete description of intertemporal distortions that are consistent with Pareto optimality in the private information case. Interestingly, we find that both negative and positive intertemporal distortions are consistent with Pareto optimality.

APPENDIX

Proof of Lemma 1

Note that removing the IC constraints (13) and (14) from the Second Best planning problem gives us exactly the First Best planning problem. Thus, neither of the two IC constraints binds at a solution to the Second Best planning problem with a given $\gamma \in [0, +\infty]$ if and only if the solution to the First Best planning problem, $c^*(\gamma)$, satisfies both IC constraints. We now show that this is the case if and only if $\gamma \in [\gamma_1, \gamma_2]$.

[17] See Kocherlakota (2006) for a survey of recent articles studying these problems. In footnote 1, we mention other relevant applications and give further references.

Substituting the expression for the First Best optimum $c^*(\gamma)$ from (8)–(11) into the IC constraint for the H-type, (15), we get

$$\frac{2\gamma}{1+\gamma} \geq \frac{2(2-\frac{10\gamma}{1+5\gamma})^5}{(\frac{10\gamma}{1+5\gamma})^5 + (2-\frac{10\gamma}{1+5\gamma})^5}.$$

Solving for γ, we get

$$\gamma \geq 5^{-\frac{5}{6}}. \tag{27}$$

This means that the First Best optimal allocation $c^*(\gamma)$ satisfies the IC condition of the H-type if and only if $\gamma \geq 5^{-\frac{5}{6}} = \gamma_1$. Similarly, substituting $c^*(\gamma)$ into the IC constraint for the L-types, expressed as in (16), and solving for γ we get

$$\gamma \leq 5^{-\frac{1}{2}}.$$

Thus, the First Best optimum $c^*(\gamma)$ satisfies the IC condition of the L-type if and only if $\gamma \leq 5^{-\frac{1}{2}} = \gamma_2$. Furthermore, the First Best optimum $c^*(\gamma)$ satisfies both IC constraints if and only if $\gamma \in [\gamma_1, \gamma_2]$.

Therefore, no IC constraints bind in the Second Best planning problem if and only if $\gamma \in [\gamma_1, \gamma_2]$. Thus, at least one IC constraint binds in the Second Best planning problem for each $\gamma \notin [\gamma_1, \gamma_2]$. We now show that exactly one IC constraint binds in this problem for each $\gamma \notin [\gamma_1, \gamma_2]$.

Suppose to the contrary that both IC constraints bind at the solution to the Second Best planning problem for some γ. Then, (i) by complementary slackness conditions, both IC constraints must be satisfied as equalities, and (ii) the solution to the Second Best planning problem for this value of γ (as for all other values) must be a Second Best Pareto optimum. Using the fact that the RF constraints hold as equalities at any solution to the Second Best planning problem (which follows from the fact that the RF constraints always bind in this problem), it is easy to check (by simply solving the RF and IC constraints for c) that both IC constraints are satisfied as equalities at only one allocation: $c = (1, 1, 1, 1)$. But this allocation is not a Second Best Pareto optimum, because an allocation $c_\varepsilon = (1+\varepsilon, 1-\varepsilon, 1-\varepsilon, 1+\varepsilon)$ Pareto-dominates c for any $\varepsilon > 0$, as the H-type strictly prefers c_ε over c, and the L-type is indifferent. (It is straightforward to confirm that c_ε is incentive compatible for ε small enough.) Thus, (i) and (ii) are inconsistent—we have a contradiction—so both IC conditions cannot bind at a solution to the Second Best planning problem for any γ.

Thus, for each $\gamma \notin [\gamma_1, \gamma_2]$ exactly one IC constraint binds in the Second Best planning problem.

Suppose now that for some $\bar{\gamma} > \gamma_2$, the IC constraint for the L-type does not bind at a solution to the Second Best planning problem, and consider a relaxed planning problem obtained from the Second Best planning problem by dropping the IC constraint of the L-type. Since this IC constraint does not

bind in the Second Best planning problem, the solution to the relaxed problem coincides with the solution to the Second Best planning problem. We know from (27) that for all $\gamma \geq \gamma_1$ the First Best optimal allocation $c^*(\gamma)$ satisfies the IC condition of the H-type. Thus, since $\bar{\gamma} > \gamma_2 > \gamma_1$, the First Best optimal allocation $c^*(\bar{\gamma})$ solves the relaxed planning problem. But then $c^*(\bar{\gamma})$ must also be the solution to the Second Best planning problem, which we know it is not, because $\bar{\gamma} \notin [\gamma_1, \gamma_2]$, a contradiction. Thus, the IC constraint of the L-type must bind in the Second Best planning problem for all $\gamma > \gamma_1$.

Similarly, supposing that the IC constraint for the H-type does not bind at a solution to the Second Best planning problem for some $\bar{\gamma} < \gamma_1$, we construct a relaxed planning problem by dropping this constraint from the Second Best planning problem, which leads to a false conclusion that $c^*(\bar{\gamma})$ solves the Second Best planning problem for a $\bar{\gamma} < \gamma_1$, a contraction. Thus, the IC constraint for the H-type must bind at a solution to the Second Best planning problem for all $\gamma < \gamma_1$.

REFERENCES

Albanesi, Stefania, and Christopher Sleet. 2006. "Dynamic Optimal Taxation with Private Information." *Review of Economic Studies* 73 (1): 1–30.

Athey, Susan, Andrew Atkeson, and Patrick J. Kehoe. 2005. "The Optimal Degree of Discretion in Monetary Policy." *Econometrica* 73 (5): 1431–75.

Atkeson, Andrew, and Robert E. Lucas, Jr. 1995. "Efficiency and Equality in a Simple Model of Efficient Unemployment Insurance." *Journal of Economic Theory* 66 (1): 64–88.

Bernanke, Ben, and Mark Gertler. 1989. "Agency Costs, Net Worth, and Business Fluctuations," *American Economic Review* 79 (1): 14–31.

Cole, Harold Lindh. 1989. "Comment: General Competitive Analysis in an Economy with Private Information." *International Economic Review* 30 (1): 249–52.

Diamond, Peter A., and James A. Mirrlees. 1978. "A Model of Social Insurance with Variable Retirement." *Journal of Public Economics* 10 (3): 295–336.

Farhi, Emmanuel, and Ivan Werning. 2006. "Progressive Estate Taxation." MIT Working Paper.

Golosov, Mikhail, Narayana Kocherlakota, and Aleh Tsyvinski. 2003.

"Optimal Indirect and Capital Taxation." *Review of Economic Studies* 70 (3): 569–87.

Golosov, Mikhail, and Aleh Tsyvinski. 2006. "Designing Optimal Disability Insurance: A Case for Asset Testing." *Journal of Political Economy* 114 (2): 257–79.

Green, Edward J., and Ping Lin. 2003. "Implementing Efficient Allocations in a Model of Financial Intermediation." *Journal of Economic Theory* 109 (1): 1–23.

Grochulski, Borys, and Narayana Kocherlakota. 2007. "Nonseparable Preferences and Optimal Social Security Systems." Minnesota Economics Research Report 2007-01.

Harris, Milton, and Robert M. Townsend. 1981. "Resource Allocation Under Asymmetric Information." *Econometrica* 49 (1): 33–64.

Hopenhayn, Hugo, and Juan Pablo Nicollini. 1997. "Optimal Unemployment Insurance." *Journal of Political Economy* 105 (2): 412–38.

Kehoe, Timothy J., David K. Levine, and Edward C. Prescott. 2002. "Lotteries, Sunspots, and Incentive Constraints." *Journal of Economic Theory* 107 (1): 39–69.

Kocherlakota, Narayana. 2005. "Zero Expected Wealth Taxes: A Mirrlees Approach to Dynamic Optimal Taxation." *Econometrica* 73 (5): 1587–621.

Kocherlakota, Narayana. 2006. "Advances in Dynamic Optimal Taxation." In *Advances in Economics and Econometrics: Theory and Applications*, Ninth World Congress, Vol. I.

Kocherlakota, Narayana, and Luigi Pistaferri. 2008. "Household Heterogeneity and Asset Trade: Resolving the Equity Premium Puzzle in Three Countries." Stanford University Working Paper.

Lucas, Robert E., Jr. 1978. "Asset Prices in an Exchange Economy." *Econometrica* 46 (6): 1429–45.

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*. New York, N.Y.: Oxford University Press.

Pavoni, Nicola, and Giovanni L. Violante. 2007. "Optimal Welfare-to-Work Programs." *Review of Economic Studies* 74 (1): 283–318.

Rogerson, William P. 1985. "Repeated Moral Hazard." *Econometrica* 53 (1): 69–76.

Stiglitz, Joseph E., and Jungyoll Yun. 2005. "Integration of Unemployment Insurance with Retirement Insurance." *Journal of Public Economics* 89 (11–12): 2037–67.