Federal Reserve Bank
of Chicago

Second Quarter 2003

Economic perspectives

Economic perspectives

Economic Perspectives is published by the Research Department of the Federal Reserve Bank of Chicago. The views expressed are the authors' and do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve System.

Single-copy subscriptions are available free of charge. Please send requests for single- and multiple-copy subscriptions, back issues, and address changes to the Public Information Center, Federal Reserve Bank of Chicago, P.O. Box 834, Chicago, Illinois 60690-0834, telephone 312-322-5111 or fax 312-322-5515.

Economic Perspectives and other Bank publications are available on the World Wide Web at http://www.chicagofed.org.

chicagofed.org

ISSN 0164-0682

Contents

Second Quarter 2003, Volume XXVII, Issue 2

Employment subcenters in Chicago: Past, present, and future

Daniel P. McMillen

Introduction and summary

Employment in large American metropolitan areas has become increasingly decentralized over time. However, employment is not distributed evenly throughout the suburban landscape. Firms congregate at highway interchanges, along rail lines, and in former satellite cities. An employment *subcenter* is a concentration of firms large enough to have significant effects on the overall spatial distribution of population, employment, and land prices. Large subcenters can look remarkably similar to a traditional central business district (CBD), with thousands of workers employed in a wide variety of industries. A *polycentric* city—a metropolitan area with a strong central business district and large subcenters—can potentially combine the advantages of the traditional centralized city and a more decentralized spatial form. Large subcenters offer agglomeration economies to firms, while potentially reducing commuting times for suburban workers. As traffic congestion increases in the suburbs, an important advantage of subcenters over more scattered employment is they can potentially be served effectively with public transportation. As a result, the location and growth patterns of subcenters in major cities are of interest to policymakers.

In this article, I document the growth of employment subcenters in the Chicago metropolitan area from 1970 to 2000. I also use employment forecasts generated by the Northeastern Illinois Planning Commission to identify subcenters for 2020. Chicago had nine subcenters in 1970. The number of subcenters rose to 13 in 1980, 15 in 1990, and 32 in 2000, and is projected to drop to 24 in 2020. Existing subcenters are becoming larger and are particularly likely to expand along major expressways. I use a formal *cluster analysis* to categorize the subcenters by employment mix in 1980, 1990, and 2000. Although Chicago's subcenters had high concentrations of manufacturing jobs in the past, the industry mix now closely resembles that of the overall metropolitan area.

I use distance from the nearest subcenter as an explanatory variable in employment and population density regressions (density is the number of workers or residents per acre). The results imply that the traditional city center still has a significant and widespread influence on densities in the Chicago metropolitan area. Firms tend to locate near important parts of the transportation system—near highway interchanges and rail stations and along freight rail lines. Subcenters also have pronounced effects on the distribution of jobs: Employment density rises significantly near subcenters. However, apart from O'Hare Airport, Chicago's subcenters are still not large enough to increase population density in neighboring areas. Construction of high-density housing near subcenters could potentially reduce aggregate commuting costs.

Subcenters are not unique to the Chicago metropolitan area. In related work, McMillen and Smith (2004) have identified subcenters in 62 large American urban areas in 1990. All but 14 of these cities have employment centers. The Los Angeles and New York metropolitan areas have the most subcenters, with 46 in Los Angeles and 38 in New York. In all 62 of these urban areas, employment density continues to decline significantly with distance from the traditional city center. Employment density also declines significantly with distance from the nearest subcenter in those cities following a polycentric form. Using the subcenter count as the dependent variable for a Poisson regression, I find that the number of subcenters rises with the urban area's population, and cities with higher commuting costs tend to have more subcenters.

Daniel P. McMillen is a professor of economics at the University of Illinois at Chicago and a consultant to the Federal Reserve Bank of Chicago.

Subcenters in the Chicago metro area

Subcenters are areas outside the traditional central business district with employment levels large enough to have significant effects on the overall spatial distribution of jobs and population. Subcenter locations are not always obvious or easy to identify beforehand. Areas near the city center with high employment density may not differ significantly from surrounding sites. Remote sites with relatively high employment densities may not have significant effects on the spatial distribution of jobs and population. Researchers such as McDonald (1987), Giuliano and Small (1991), Craig and Ng (2001), and McMillen (2001) have proposed procedures that objectively identify subcenter sites using standard data sources.

In this article, I use Giuliano and Small's (1991) approach to identify subcenters in the Chicago metropolitan area between 1970 and 2000 and to predict subcenter sites in 2020. Analyzing the Los Angeles metropolitan area, Giuliano and Small define a subcenter as a set of contiguous tracts that each have at least ten employees per acre and together have at least 10,000 employees.[1] The number of subcenters is sensitive to these two cutoffs. Higher minimum density levels or higher values for total employment produce fewer subcenters. To ensure reasonable results, one needs local knowledge to guide the choice of cutoffs. After some experimentation, I chose cutoff points of 15 employees per acre and 10,000 total workers. These values produce a reasonable number of subcenters in each period. McMillen and Smith (2004) provide a detailed explanation of the subcenter identification procedure.

Data on employment and population were provided by the Northeastern Illinois Planning Commission (NIPC). NIPC conducts decennial land use surveys for the six-county Chicago primary metropolitan statistical area. The six counties are Cook, DuPage, Kane, Lake, McHenry, and Will. The unit of observation is the quarter section, which is 160 acres or one-quarter of a square mile. There are slightly more than 15,000 quarter sections in these six counties. NIPC provided employment data for 1970, 1980, 1990, and 2000, and forecasts for 2020. Population data are not yet available for quarter sections in 2000, although forecasts are available for 2020. Comparisons over time for individual quarter sections are not completely reliable because NIPC has changed its methodology. In 1970 and 2020, NIPC reports employment data for any quarter section with jobs. In 1980 and 1990, only quarter sections with ten or more employees are included in the dataset, whereas the minimum employment level is eight in 2000. Due to this limitation, the dataset has more tracts with positive values for employment in 1970 than in 1980–2000, despite the general decentralization of the Chicago metropolitan area over this time.

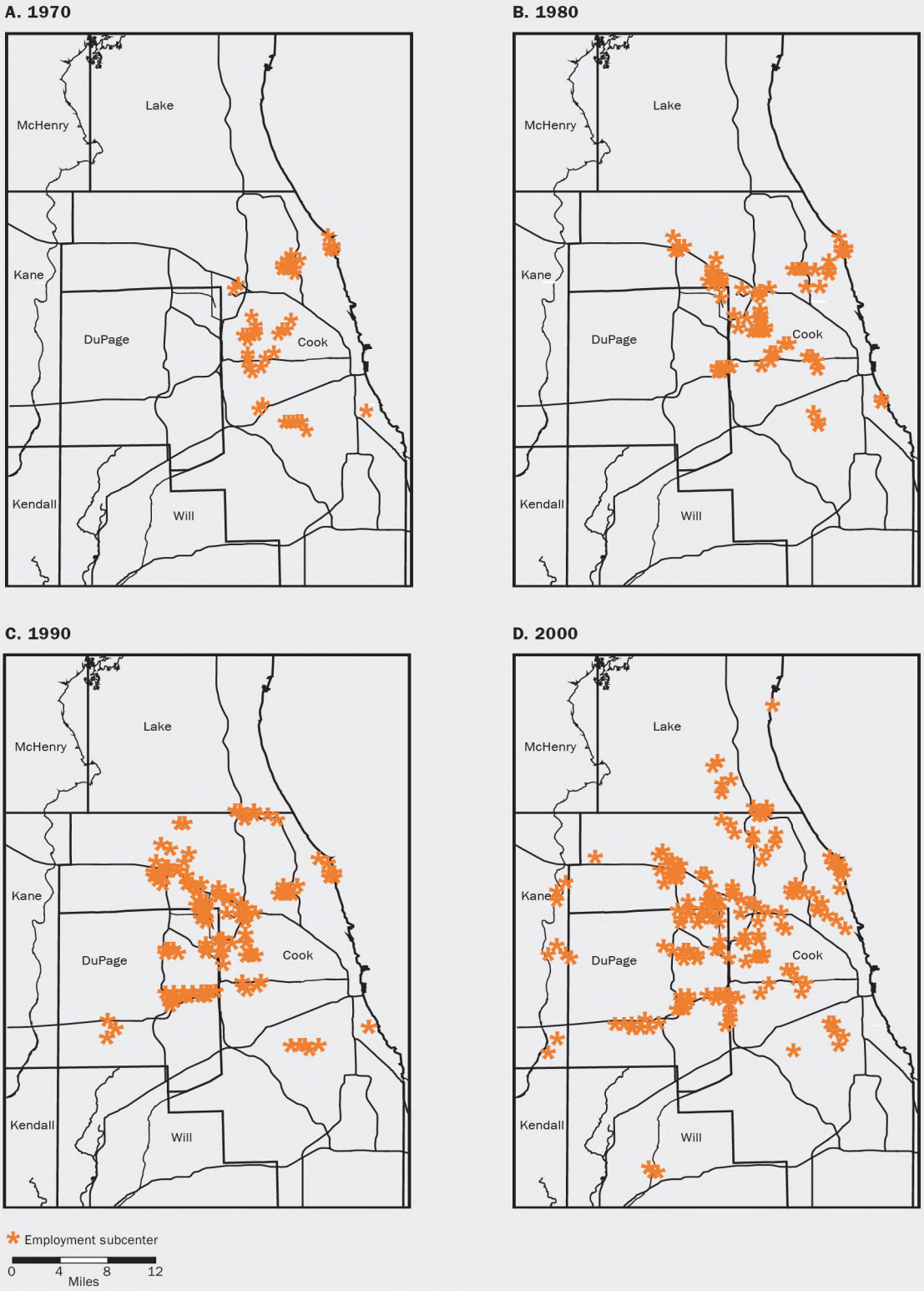
Figures 1 and 2 show the subcenter sites. The number of subcenters rises from nine in 1970 to 13 in 1980, 15 in 1990, and 32 in 2000. The NIPC employment forecasts lead to a prediction of 24 subcenters in 2020. Figure 1, panel A shows that in 1970 there was a subcenter in Hyde Park on the south side of Chicago, along with a ring of subcenters that nearly encircles the city. The number and geographic scope of the subcenters expand over time. O'Hare Airport is the center of a large conglomeration of subcenter employment. Another group of subcenters spreads along the I-88 toll way running west out of the city. In 2000 (panel D), small subcenters appear at the fringes of the metropolitan area in Kane County and Will County. These sites are in the old satellite cities of Elgin, St. Charles, Aurora, and Joliet. The NIPC forecasts suggest that the satellite cities will not continue to qualify for subcenter status in 2020, although the accuracy of this forecast appears questionable in light of the ongoing decentralization of employment in the Chicago metropolitan area. In 2020, also, several formerly separate subcenters along I-88 and near O'Hare are predicted to merge (figure 2). The general pattern of figure 1 is one of rapidly expanding subcenters, with most of the growth occurring near O'Hare Airport and along the major highways serving the city.

Subcenter clusters

Employment data are available by sector for 1980, 1990, and 2000. Table 1 presents data on the total number of jobs and the distribution of employment across five sectors in the subcenters identified for these years. The sectors are manufacturing; retail; services; transportation, communication, and utilities (TCU); finance, insurance, and real estate (FIRE); and government (federal, state, and local). I also use these sectors as headings for groups of similar subcenters that I identify using a formal cluster analysis. The cluster analysis[2] categorizes subcenters by looking for groups with similar employment compositions. The cluster analysis is performed for a given number of clusters, leaving it to the analyst to specify the appropriate number. Experimentation suggested that specifying five groups produces reasonable results, with clusters that are dominated by jobs in one of the five primary employment categories. Table 1 groups the subcenters by cluster in each year, with the subcenter sites identified by the municipalities (or neighborhoods within Chicago) in which they are located.

FIGURE 1

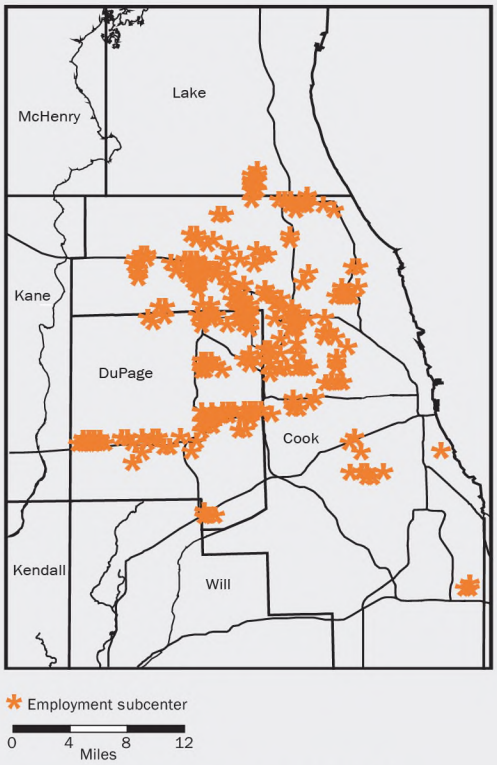
Subcenter locations



Source: Author's calculations based upon data from the Northeastern Illinois Planning Commission, 1970–2000, decennial land use surveys.

FIGURE 2

Subcenter locations in 2020



Source: Author's calculations based upon data from the Northeastern Illinois Planning Commission, 1970–2000, decennial land use surveys.

In 1980, eight of the 13 subcenters were dominated by manufacturing jobs. Traditional manufacturing sites such as Cicero, the Clearing District of Chicago, and Franklin Park appear as subcenters, along with newer suburban sites such as Elk Grove Village, Niles–Skokie, and Schaumburg. The manufacturing subcenters are generally larger than the service, TCU, and government subcenters, with total employment ranging from 13,430 in Rosemont to 46,740 in Franklin Park–Melrose Park. Although these subcenters are dominated by manufacturing, they also can include significant numbers of other types of jobs. For example, 28.89 percent of the Albany Park subcenter's jobs are in the FIRE sector, compared with 55.05 percent in manufacturing. The Clearing–West Lawn and Schaumburg subcenters have many retail jobs, representing 27.73 percent and 32.04 percent of the jobs in those subcenters, respectively. Rosemont is a diversified subcenter, having a similar number of jobs in manufacturing, service, TCU, and FIRE. Of the remaining subcenters in 1980, three specialize in the service sector (Evanston, Oak Brook, and the Hyde Park area of Chicago, which includes the University of Chicago), one specializes in TCU (O'Hare), and one specializes in government (Broadview–Maywood–Oak Park). Maywood has a significant county governmental facility, Broadview has several township offices, and Oak Park, which is fairly large in population, has several village and township offices. Oak Brook, which is the site of a regional shopping mall and is near the intersection of the Tri-State and East–West tollways, also includes many retail and TCU jobs: These two sectors account for 22.95 percent and 21.78 percent of the jobs in the subcenter, respectively.

Table 1 shows that the subcenters continue to be dominated by manufacturing jobs in 1990, although the locations have changed somewhat. Whereas the manufacturing subcenters were formerly concentrated in Chicago and in the near western suburbs, by 1990 they are more apt to be in the northwestern suburbs and near O'Hare Airport. New manufacturing sites in this area include Addison, Arlington Heights, and Palatine. Another new manufacturing subcenter appears in the rapidly growing western suburb of Naperville. These manufacturing subcenters range in size from 10,120 in Naperville to 95,420 in Elk Grove Village–Schaumburg. Several of the subcenters also include many TCU jobs, although they are placed in another category: TCU accounts for 38.51 percent of the jobs in the Addison subcenter, 20.21 percent in Bedford Park–Chicago Lawn–West Lawn, 22.97 percent in Des Plaines–Rosemont, 27.10 percent in Elk Grove Village–Schaumburg, 24.90 percent in Naperville, 21.65 percent in Niles–Skokie, and 40.82 percent in Palatine. Five subcenters specialize in service employment in 1990: The sector accounts for 48.78 percent of the employment in Bellwood–Broadview–Maywood, 32.47 percent in Deerfield–Northbrook, 57.74 percent in Evanston, 39.40 percent in Oak Brook, and 98.80 percent at the University of Chicago. The O'Hare subcenter continues to be dominated by TCU employment in 1990. None of the subcenters is placed in the government category in 1990.

The list grows to 32 subcenters in 2000 from 15 in 1990. The number of manufacturing subcenters falls to six—Addison, Glenview, North Chicago, Schaumburg, St. Charles, and Wheeling. All the manufacturing subcenters are now in more distant suburbs. Retail appears as a subcenter category in 2000, with sites in Deerfield–Northbrook (classified as service in 1990), Franklin Park, Hoffman Estates, and Melrose Park. The Hoffman Estates subcenter is a result of the movement of the Sears corporate headquarters out of Chicago. The number of service sector subcenters also increases significantly, with sites in Aurora, Broadview–Forest Park, Cicero–Oak Park, Elk Grove Village,

TABLE 1

Subcenter characteristics

Subcenter	Cluster	Total employment	Subcenter employment composition (%) Mfg.	Retail	Services	TCU	FIRE	Government
			1980					
Albany Park–Jefferson Park–North Park	mfg	14,640	55.05	0.41	2.53	3.48	28.89	6.63
Cicero–Austin	mfg	28,210	62.96	17.23	4.86	8.29	0.00	2.30
Clearing–West Lawn	mfg	10,890	45.36	27.73	1.65	10.65	3.86	10.74
Elk Grove Village	mfg	37,030	39.08	7.05	3.83	44.56	0.16	0.00
Franklin Park–Melrose Park	mfg	46,740	68.66	11.49	2.55	9.52	0.06	5.88
Niles–Skokie	mfg	40,800	65.66	5.96	5.27	17.28	1.81	0.20
Rosemont	mfg	13,430	18.63	8.59	24.11	25.89	17.96	1.55
Schaumburg	mfg	23,000	46.22	32.04	4.13	6.61	6.00	4.91
Evanston	serv	22,430	3.79	12.26	51.67	2.41	25.28	4.15
Oak Brook	serv	27,500	12.36	22.95	30.25	21.78	10.40	0.55
University of Chicago (Hyde Park)	serv	15,300	0.07	0.26	96.08	0.33	0.07	2.75
O'Hare	tcu	11,970	0.00	20.55	10.69	51.04	0.00	17.71
Broadview–Maywood–Oak Park	govt	22,260	7.23	10.42	36.21	12.62	5.75	27.18
			1990					
Addison	mfg	11,790	42.32	0.85	7.46	38.51	0.34	1.10
Arlington Heights	mfg	15,270	55.73	4.98	7.60	7.99	4.58	0.00
Bedford Park–Chicago Lawn–West Lawn	mfg	16,230	49.23	5.67	12.14	20.21	0.25	12.26
Des Plaines–Rosemont	mfg	44,070	24.95	8.46	25.86	22.97	10.29	3.53
Elk Grove Village–Schaumburg	mfg	95,420	33.03	11.04	15.05	27.10	7.68	1.07
Elmhurst–Franklin Park–Melrose Park–Northlake	mfg	50,250	46.61	14.31	13.47	14.83	1.59	6.79
Naperville	mfg	10,120	17.00	13.54	22.73	24.90	2.87	12.45
Niles–Skokie	mfg	27,620	45.08	5.25	14.45	21.65	5.54	0.91
Palatine	mfg	10,290	49.17	2.82	4.76	40.82	2.04	0.00
Bellwood–Broadview–Maywood	serv	21,730	17.67	1.47	48.78	12.43	0.18	17.35
Deerfield–Northbrook	serv	26,730	17.92	22.15	32.47	16.46	7.00	0.19
Evanston	serv	25,580	7.00	12.51	57.74	2.15	13.88	3.91
Oak Brook	serv	76,760	7.43	18.88	39.40	20.64	8.91	0.81
University of Chicago (Hyde Park)	serv	16,670	0.00	0.54	98.80	0.12	0.00	0.00
O'Hare	tcu	40,340	0.00	9.22	7.68	76.03	0.05	6.79
			2000					
Addison	mfg	29,593	33.12	8.01	10.57	38.03	0.56	0.00
Glenview	mfg	15,215	40.47	5.49	24.96	23.35	0.20	0.00
North Chicago	mfg	19,432	88.30	0.00	0.00	11.70	0.00	0.00
Schaumburg	mfg	82,092	40.01	13.11	19.39	6.80	3.55	0.00
St. Charles	mfg	10,815	51.20	16.98	16.38	6.03	2.64	4.28
Wheeling	mfg	10,595	24.68	1.52	25.16	16.64	0.28	0.00
Deerfield–Northbrook	retl	51,253	4.06	49.45	23.79	14.07	3.80	1.50
Franklin Park	retl	25,064	30.93	47.12	2.29	16.66	0.00	0.37
Hoffman Estates	retl	17,355	0.00	100.00	0.00	0.00	0.00	0.00
Melrose Park	retl	54,550	6.19	71.37	16.18	4.41	0.92	0.00
Aurora–South	serv	10,570	0.52	1.96	50.23	1.42	13.52	19.01
Broadview–Forest Park	serv	28,119	8.70	0.79	88.33	1.64	0.00	0.00
Cicero–Oak Park	serv	15,609	3.57	5.45	63.58	3.13	8.50	2.94
Elk Grove Village	serv	101,012	20.92	4.19	36.98	22.73	9.88	1.11
Evanston	serv	46,957	1.08	5.40	72.00	0.37	1.09	14.01
Glenbard	serv	28,242	3.83	16.07	57.40	13.26	8.31	0.00
Joliet	serv	10,917	0.35	5.06	43.67	2.83	4.09	21.89
Lincolnshire	serv	33,121	5.64	3.34	78.27	11.85	0.00	0.00
Lisle–Naperville	serv	34,197	8.76	14.42	40.79	17.02	16.16	0.66
Oak Brook	serv	78,810	3.56	19.17	49.74	12.15	13.24	0.05
Bedford Park	tcu	18,790	4.28	0.00	1.37	94.10	0.00	0.00
Bensenville–Elmhurst	tcu	29,253	17.71	9.65	28.82	37.45	2.22	0.00
Midway Airport	tcu	20,183	12.35	15.40	14.16	35.18	22.22	0.00
O'Hare	tcu	61,527	0.00	9.57	3.02	87.37	0.00	0.00
Vernon Hills	tcu	13,599	11.42	9.25	8.93	45.97	24.42	0.00
Prospect Heights	fire	20,913	4.19	6.16	1.32	2.55	85.77	0.00
Arlington Heights	govt	14,270	5.88	1.97	23.75	5.05	1.85	30.07
Aurora–North	govt	14,268	0.00	0.00	0.00	0.00	0.00	99.41
Des Plaines–Rosemont	govt	67,565	19.27	2.82	28.33	16.19	4.72	27.69
Elgin	govt	26,119	11.58	0.89	12.68	0.50	0.00	61.50
Niles–Skokie–Northern Chicago	govt	59,806	30.68	4.03	23.16	8.18	1.83	27.63
Norridge–Norwood Park	govt	16,662	16.20	1.75	31.53	2.04	0.18	46.61

Notes: Mfg. is manufacturing; TCU is transportation, communications, and utilities; and FIRE is finance, insurance, and real estate.
Source: Northeastern Illinois Planning Commission, 1970–2000, decennial land use surveys.

Evanston, Glenbard, Joliet, Lincolnshire, Lisle–Naperville, and Oak Brook. In addition, TCU accounts for five subcenters in 2000, one subcenter specializes in FIRE, and six have large concentrations of government employment. The largest subcenters are in Schaumburg (82,092 employees) and Elk Grove Village (101,012 employees). In 2000, the subcenter job mix closely resembles the employment composition of the full metropolitan area.[3]

Employment and population density in Chicago

The spatial distribution of jobs and residences can be summarized by regressing measures of employment and population density on a set of explanatory variables, including distance from Chicago's traditional city center and measures of proximity to subcenter sites. Population density functions have a long history in urban economics, dating back to Clark (1951). Issues involved in estimation and a review of studies up to the late 1980s are reviewed in McDonald (1989). Employment density functions are estimated less frequently. Prominent examples include Booth (1999), Combes (2000), McDonald (1985), McDonald and Prather (1994), McMillen and McDonald (1997), and Small and Song (1994). With the natural logarithm of density as the dependent variable, the coefficient for distance from the central business district (CBD) or city center is referred to as the "CBD gradient." The gradient measures the percentage change in density associated with a one-mile increase in distance from the city center. It is a simple measure of centralization: Density declines rapidly with distance in a highly centralized city, leading to large negative values for the estimated CBD gradient. Empirical studies suggest that most cities in the world have become increasingly decentralized over the last century, although employment generally remains more centralized than population.

Explanatory variables for the estimated density functions include distance from the traditional city center at the intersection of State and Madison streets, distance from O'Hare Airport, and distance from the nearest quarter section that is part of a subcenter. Distance from the nearest subcenter enters the estimating equations in inverse form, because I expect the effect of proximity to a subcenter to decline rapidly with distance. Proximity to subcenters increases densities if the coefficient for this variable is positive, and the effect rises over time if the coefficient becomes larger over time.

Other explanatory variables have localized effects on densities that can be accounted for using simple dummy variables. I include dummy variables that equal one when a quarter section is within one-third of a mile and between one-third and one mile of the following sites: a highway interchange, a commuter rail station, an elevated train line (the "el"), a station on an electric line serving the South Side, and Lake Michigan. I distinguish between commuter rail, el, and electric train lines because they have different areas and clienteles. The commuter rail lines primarily serve the suburbs, and have long intervals between stops. El lines are nearly entirely within the City of Chicago, and have frequent stops. The electric train line is something of a hybrid. It runs from downtown Chicago to the distant southern suburbs, along with a separate spur to Northwest Indiana. Although it primarily serves suburbanites, it resembles the el in making frequent stops within the city.

Table 2 presents detailed employment density estimates. The results indicate that employment fell by 5.6 percent with each mile from the Chicago city center in 1970. The rate of decline falls to 2.2 percent in 1980 as Chicago becomes more decentralized, and remains at about that level for 1990 (2.3 percent) and 2000 (2.2 percent again). The rate of decline is expected to be 1.9 percent per mile in 2020, based on NIPC employment forecasts. With the exception of 2000, proximity to O'Hare also increases employment density. Employment density is estimated to decline by 1.0 percent per mile in 1980, 0.9 percent in 1990, and a forecasted 3.4 percent in 2020.

Other results in table 2 are much as expected. Employment density is higher near highway interchanges. Densities are estimated to be 30.6 percent higher within one-third of a mile of a highway interchange in 1970, compared with 37.9 percent in 2000, and a forecasted 40.5 percent in 2020. Densities decline somewhat in the next two-thirds of a mile from a highway interchange. In 1970, densities are 18.1 percent higher in the ring from one-third to one mile of a highway interchange than in more distant sites, compared with 21.6 percent in 2000 and a forecasted 13.6 percent in 2020. Similarly, densities are higher near commuter rail stations. For example, in 1970 employment density is estimated to be 85.2 percent higher within one-third of a mile of a commuter station and 50.6 percent higher in the one-third to one-mile ring, compared with more distant locations. Commuter train stations decline in importance in subsequent years. In 2020, employment density is expected to be 54.5 percent higher within one-third of a mile of a commuter station and 9.4 percent higher in the one-third to one-mile ring. Proximity to stations on the electric line has similar effects on employment, except the effect is confined to the initial zero to one-third of a mile ring.

TABLE 2

Total employment density

	1970	1980	1990	2000	2020
Miles from city center	–0.056	–0.022*	–0.023*	–0.022*	–0.019*
	(15.742)*	(7.685)	(9.311)	(6.334)	(7.448)
Miles from O'Hare Airport	–0.005	–0.010*	–0.009*	0.005	–0.034*
	(1.599)	(3.521)	(3.461)	(1.423)	13.523)
0 – 1/3 mile from highway interchange	0.306*	0.266*	0.287*	0.379*	0.405*
	(3.132)	(3.759)	(4.478)	(5.048)	(5.678)
1/3 – 1 mile from highway interchange	0.181*	0.259*	0.180*	0.216*	0.136*
	(2.996)	(5.682)	(4.309)	(4.514)	(2.974)
0 – 1/3 mile from commuter rail station	0.852*	0.541*	0.576*	0.608*	0.545*
	(5.858)	(5.102)	(5.632)	(5.141)	(4.619)
1/3 – 1 mile from commuter rail station	0.506*	0.182*	0.106*	0.127*	0.094**
	(8.123)	(3.839)	(2.414)	(2.520)	(1.891)
0 – 1/3 mile from el station	0.937*	0.770*	1.038*	0.551*	1.152*
	(5.752)	(6.401)	(9.147)	(4.232)	(8.634)
1/3 – 1 mile from el station	0.557*	0.291*	0.592*	0.146	0.500*
	(4.877)	(3.400)	(7.339)	(1.558)	(5.349)
0 – 1/3 mile from station on electric line	0.805*	0.553*	0.572*	0.564*	0.770*
	(2.887)	(2.711)	(2.952)	(2.560)	(3.431)
1/3 – 1 mile from station on electric line	0.173	0.173**	–0.036	0.027	0.329*
	(1.244)	(1.671)	(0.376)	(0.233)	(3.012)
0 – 1/3 mile from Lake Michigan	–0.207	0.015	–0.228	0.197	0.173
	(1.054)	(0.090)	(1.478)	(1.065)	(0.991)
1/3 – 1 mile from Lake Michigan	0.276*	0.223*	0.005	0.100	0.267*
	(2.019)	(2.084)	(0.049)	(0.863)	(2.307)
Chicago River or canal runs through tract	0.386	0.433*	0.284**	0.583*	–0.020
	(1.552)	(2.532)	(1.719)	(2.906)	(0.108)
Freight rail line within tract	0.723*	0.398*	0.356*	0.250*	0.430*
	(12.893)	(9.305)	(9.122)	(5.546)	(10.105)
Within City of Chicago	1.035*	0.396*	0.135*	–0.044	0.065
	(11.917)	(5.909)	(2.174)	(0.601)	(0.936)
Inverse of distance from the nearest subcenter	0.774*				0.568*
	(19.209)				(27.253)
Nearest subcenter is in retail cluster				0.044	
				(0.448)	
Nearest subcenter is in government cluster		0.164		0.208*	
		(1.123)		(2.195)	
Nearest subcenter is in service cluster		–0.037	0.047	–0.209*	
		(0.823)	(1.001)	(2.746)	
Nearest subcenter is in TCU cluster		0.431	–4.978*	–0.141	
		(1.200)	(1.965)	(1.642)	
Nearest subcenter is in FIRE cluster				–1.218*	
				(2.390)	
Inverse of distance from nearest subcenter × retail cluster				0.758*	
				(13.840)	
Inverse of distance from nearest subcenter × government cluster		0.471*		0.645*	
		(5.161)		(13.685)	
Inverse of distance from nearest subcenter × service cluster		0.564*	0.600*	0.784*	
		(11.493)	(19.102)	(24.082)	
Inverse of distance from nearest subcenter × TCU cluster		0.754*	2.510*	0.756*	
		(4.016)	(3.707)	(11.237)	
Inverse of distance from nearest subcenter × manufacturing cluster		0.665*	0.670*	0.768*	
		(23.708)	(28.658)	(17.003)	
Inverse of distance from nearest subcenter × FIRE cluster			0.992*		
				(3.707)	
Constant	–0.325*	0.253*	0.403*	0.078*	0.840*
	(3.553)	(3.345)	(5.677)	(0.799)	(11.308)
R^2	0.425	0.364	0.390	0.314	0.376
Number of observations	6,081	5,220	5,817	5,649	7,522

Notes: Absolute *t*–values are in parentheses. The dependent variable is the natural logarithm of employment density per acre. "*" indicates significance at the 5 percent level; and "**" indicates significance at the 10 percent level.

Source: Author's calculations based on data from the Northeastern Illinois Planning Commission, 1970–2000, decennial land use surveys.

Lake Michigan has little or no effect on employment density. Quarter sections through which the Chicago River or the Sanitary and Ship Canal runs tend to have high employment density. In 2000, densities are estimated to be 58.3 percent higher in quarter sections with the river or canal. Although sites within Chicago had higher densities from 1970 to 1990, the effect declines from a 103.5 percent increase in 1970 to 39.6 percent in 1980 to 13.5 in 1990. After controlling for other explanatory variables, city locations do not have higher employment density in 2000 or 2020.

The final set of results in table 2 includes the effects of proximity to subcenters on employment density. The 1970 and 2020 regressions include a single variable representing the inverse of distance from the nearest subcenter. The regressions confirm the importance of subcenters in accounting for the spatial distribution of employment density. Letting d represent the distance from the nearest subcenter, the marginal effect of distance is $-0.774/d^2$ in 1970 and a forecasted $-0.568/d^2$ in 2020. The minimum value for d is 0.25. Thus, the estimated marginal effect of distance from the nearest subcenter in 1970 is –12.38 at subcenter sites, with the effect falling to –0.77 after one mile, and –0.19 after two miles. Comparable values for 2020 are –9.09, –0.57, and –0.142, respectively. Although subcenters do not affect employment over as wide an area as the traditional CBD, the high t-values of 19.209 in 1970 and 27.253 in 2020 indicate that they are critically important determinants of the spatial distribution of jobs in the Chicago area.

For the years with data on employment sectors (1980, 1990, and 2000), I include separate explanatory variables for each cluster type. For these years, the regressions include dummy variables indicating the sector for the closest subcenter and interactions between these dummy variables and the inverse of distance from the subcenter. The dummy variables are generally not statistically significant. The coefficients for the inverse of distance from the nearest subcenter again indicate that employment densities rise significantly near subcenters. In 1980, the marginal effect of distance from the nearest subcenter is –0.471 at a distance of one mile when the nearest subcenter is in the government cluster, compared with –0.564 for service subcenters, –0.754 for TCU, and –0.665 for manufacturing. In 1980, these marginal effects are –0.600 for service, –2.510 for TCU, and –0.670 for manufacturing. In 2000, the marginal effect at one mile from a subcenter is –0.758 for retail, –0.645 for government, –0.784 for service, –0.756 for TCU, and –0.768 for manufacturing. The results are all highly significant. What is more surprising is that, with the exception of the TCU cluster in 1990, the estimated marginal effects do not vary much across sectors.

Table 3 presents abbreviated results for comparable population density function estimates. Population density is estimated to decline by 7.3 percent with each mile from the Chicago city center in 1970, compared with 7.8 percent in 1980, 7.2 percent in 1990, and a forecasted 6.6 percent in 2020 (recall that population data are not yet available for 2000 at the quarter section level). These results are somewhat surprising in their implication that the CBD gradient is now larger for population than for jobs after controlling for the effects of other variables. O'Hare Airport also has a significant effect on population density. Controlling for other variables, each additional mile from O'Hare reduces population density by 4.9 percent in 1970, 4.6 percent in 1980, 6.0 percent in 1990, and a forecasted 5.9 percent in 2020.

In keeping with the results of McMillen and McDonald (2000), proximity to employment subcenters is estimated to *reduce* population density. Each additional mile from the nearest employment subcenter increases density by 16.4 percent in 1970, 44.9 percent in 1980, 36.2 percent in 1990, and a forecasted 47.3 percent in 2002. This result has two explanations. First, our density measures are gross rather than net, meaning that density is measured per acre of total land area rather than per acre of residential land area. Densities are low near subcenters because by definition much of the land area in subcenters is in nonresidential use.

TABLE 3

Population density

	1970	1980	1990	2020
Miles from city center	–0.073 (30.963)	–0.078 (31.107)	–0.072 (29.670)	–0.066 (28.345)
Miles from O'Hare Airport	–0.049 (22.795)	–0.046 (20.052)	–0.060 (25.880)	–0.059 (25.942)
Inverse of distance from nearest subcenter	–0.164 (4.305)	–0.449 (13.263)	–0.362 (12.589)	–0.473 (18.731)
R^2	0.512	0.423	0.429	0.361
Number of observations	10369	10942	11129	11687

Notes: Absolute t-values are in parentheses. The dependent variable is the natural logarithm of population density. Other explanatory variables include dummy variables representing locations within the City of Chicago and proximity to highways, commuter rail lines, el lines, electric train lines, Lake Michigan, the Chicago River and the Sanitary and Ship Canal, and freight rail lines.

Source: Author's calculations based on data from the Northeastern Illinois Planning Commission, 1970–2000, decennial land use surveys.

Second, although subcenters are getting bigger, they are not yet large enough in the Chicago area to lead to large increases in population density in neighboring sites. Subcenter employment has increased primarily through an increase in the number of subcenters rather than by the creation of a few larger subcenters that rival the traditional CBD in their effects on density patterns.

Subcenters in other metro areas

Subcenters are not only a Chicago phenomenon. Studies by Anderson and Bogart (2001), Bogart and Ferry (1999), Cervero and Wu (1997, 1998), Craig and Ng (2001), Giuliano and Small (1991), McMillen (2001), and Small and Song (1994) have identified subcenters in Cleveland, Dallas, Houston, Indianapolis, Los Angeles, New Orleans, St. Louis, and the San Francisco Bay Area. Recently, Baumont, Ertur, and LeGallo (2002) and Muñiz, Galindo, and García (2003) have extended the analysis to the European cities of Dijon, France and Barcelona, Spain.

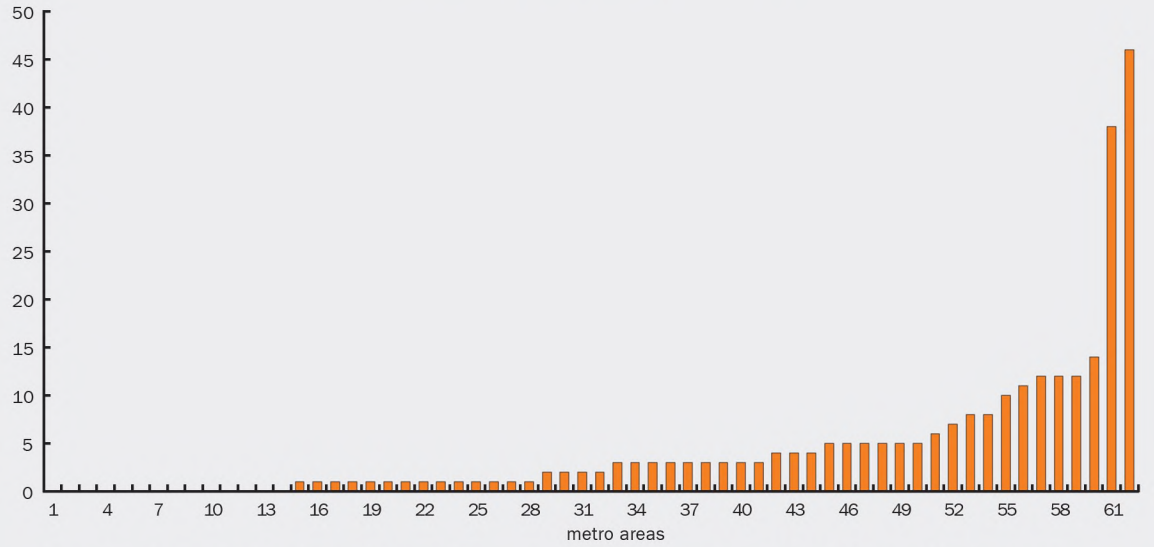
The remainder of this section summarizes the results of a recent study by McMillen and Smith (2004), which is the first to apply a single subcenter identification procedure to a large number of metropolitan areas. They use a variant of the Giuliano and Small (1991) procedure to identify subcenters in 62 large U.S. metropolitan areas. The data come from the urban element of the Census Transportation Planning Package, which is produced by the Department of Transportation's Bureau of Transportation Statistics (BTS). The BTS obtained special tabulations of 1990 U.S. Census data to match Census data with the BTS geographic unit, called the Transportation Analysis Zone. These zones, which vary across metropolitan areas, are typically smaller than Census tracts or zip codes and often are the same as Census blocks.

Figure 3 shows the distribution of the number of subcenters across the 62 metropolitan areas. Fourteen of the metropolitan areas have no subcenters. Eight metropolitan areas—Boston, Chicago, Dallas–Fort Worth, Los Angeles, New York, the San Francisco Bay Area, Seattle, and Washington, DC—have at least ten subcenters. The two largest cities, New York and Los Angeles, have the most subcenters with 38 and 46, respectively. Chicago is next with 15.

For a subset of the 62 metropolitan areas, table 4 presents the results of simple regressions of the natural logarithm of employment density on distance from the traditional central business district and the inverse of distance from the nearest zone that is part of a subcenter. The R^2s for the regression indicate that these two variables alone account for no less than 21.7 percent of the variation in employment density (in San Francisco), with an average of 38.3 percent and a maximum of 57.0 percent (in Washington, DC). The traditional CBD still has a tremendous impact on employment

FIGURE 3

Number of subcenters by metro area



Source: Author's calculations based upon data from the Northeastern Illinois Planning Commission, 1970–2000, decennial land use surveys.

TABLE 4

Employment density functions for selected metro areas, 1990

Metro area	No. of subcenters	CBD coefficient	*t*–value	Subcenter coefficient	*t*–value	R^2	*n*
Atlanta	4	–0.213	–34.307	0.482	9.266	0.569	943
Boston	11	–0.097	–31.519	0.353	19.013	0.267	3744
Chicago	15	–0.042	–28.760	0.649	36.607	0.340	5935
Cincinnati	3	–0.170	–20.370	0.297	3.656	0.313	958
Cleveland	3	–0.138	–24.385	0.199	3.153	0.399	991
Dallas	12	–0.089	–34.565	0.532	26.049	0.297	4379
Denver	5	–0.095	–19.545	0.419	8.864	0.277	1336
Detroit	8	–0.106	–35.274	0.484	19.342	0.388	2688
Houston	8	–0.118	–33.938	0.583	14.747	0.399	2128
Kansas City	2	–0.227	–26.767	0.487	5.896	0.529	732
Los Angeles	46	–0.048	–16.779	0.449	20.846	0.201	3051
Minneapolis–St. Paul	7	–0.201	–34.608	0.373	11.020	0.559	1187
New York	38	–0.097	–77.606	0.172	17.631	0.306	14831
Philadelphia	4	–0.109	–25.083	0.527	9.072	0.363	1350
Phoenix	5	–0.206	–31.049	0.308	7.923	0.545	996
San Diego	6	–0.090	–15.121	0.335	6.369	0.299	632
San Francisco	12	–0.056	–24.195	0.378	16.080	0.217	2913
Seattle	14	–0.133	–21.009	0.438	11.255	0.404	828
St. Louis	5	–0.165	–24.630	0.453	7.378	0.420	995
Washington, DC	10	–0.153	–55.863	0.416	22.782	0.570	3090

Note: The explanatory variables include an intercept, distance from the city center, and the inverse of distance from the nearest subcenter.
Source: Author's calculations based upon data from the U.S. Department of Transportation, Bureau of Transportation Statistics, *Census Transportation Planning Package*.

densities. For example, employment densities in Atlanta are estimated to decline by 21.3 percent with each additional mile from the CBD after controlling for proximity to subcenters. In table 4, the average CBD gradient is –12.8 percent, with a range of –4.2 percent in Chicago to –22.7 percent in Kansas City.

The coefficients for the inverse of distance from the nearest subcenter zone indicate that employment densities are higher near subcenters. For example, in Atlanta the estimated marginal effect of distance from the nearest subcenter is estimated to be $-0.482/d^2$, where d is distance. The marginal effect of distance is –7.71, –.48, and –.12 for sites that are one-quarter mile, one mile, and two miles, respectively, from the nearest subcenter in Atlanta. The average coefficient for distance from the nearest subcenter is 0.417 in table 4, with a range of 0.172 (New York) to 0.649 (Chicago). These results imply that the rate of decline in employment densities with distance from the nearest subcenter is highest in Chicago and lowest in New York.

Theoretical and empirical models of subcenter formation have thus far developed in relative isolation. Theoretical models have focused on examining the equilibrium spatial configuration of polycentric cities rather than on producing empirically testable, comparative static results. Models such as those developed by Anas and Kim (1996), Berliant and Konishi (2000), Fujita, Krugman, and Mori (1999), Fujita and Ogawa (1982), Fujita, Thisse, and Zenou (1997), Helsley and Sullivan (1991), Henderson and Mitra (1996), Konishi (2000), Wieand (1987), and Yinger (1992) emphasize the role that population and commuting cost play in altering the equilibrium spatial configuration of a city. The primary prediction is that the equilibrium number of subcenters tends to rise with population and commuting costs.

This prediction can be tested for our sample of 62 metropolitan areas using the number of subcenters as the dependent variable. Poisson regression is the appropriate estimation procedure for this type of count data (Cameron and Trivedi, 2001). The key explanatory variables are population and commuting costs. Population, which is measured over the full metropolitan area, ranges from 127,855 in Laredo, Texas, to 16,885,598 in New York. I use two measures of commuting cost. The first is a travel time index developed by the Texas Transportation Institute for its Mobility Monitoring Program. It is designed as a measure of peak-period congestion. The Travel Rate Index exceeds 1.0 if it takes longer on average to make a trip in congested periods than at other times of the day. As an alternative, I also use a measure of highway capacity—thousands of miles traveled on average daily by all vehicles per mile of freeway lanes (*DVMTLANE*). This index focuses on average travel time across the day, whereas the travel time index focuses on travel at peak commuting

times. Its advantage is that it has a greater claim to being exogenous or predetermined: The highway capacity in most American cities is a direct result of federal highway programs from the 1950s and 1960s. Strict exogeneity is not essential because I am estimating an equilibrium relationship. The correlation is high among all of the indexes available from the Texas Transportation Institute's Urban Mobility Study, and the results are not sensitive to the choice.

Other explanatory variables control for differences among cities. I include the central city's proportion of the urban area's population, because subcenters may be more likely to form when there are more suburbs. Competition among suburbs for firms may produce subcenters, whereas a large central city may adopt policies to encourage the continued dominance of the traditional CBD. The median income of the central city has ambiguous effects on subcenter formation. On the one hand, high income suggests a vibrant central city, which may discourage subcenter formation. But incomes in the central city and suburbs are highly correlated, and subcenters may be more likely to form if higher income increases the aversion to long commutes.

TABLE 5

Poisson regressions: Number of subcenters

	Number of subcenters		
	Model 1	Model 2	Model 3
Metro population (millions)	0.148* (7.015)	0.151* (7.670)	0.173* (12.846)
Travel Rate Index	1.223* (2.878)		
DVMTLANE		0.094 (3.190)	0.093 (5.441)
Proportion of metro population in central city	–1.479* (2.494)	–1.490* (2.522)	–1.710* (3.694)
Median income in central city ($1,000)	0.029 (1.512)	0.027 (1.409)	
Average tract size (sq. miles)	–0.053 (1.293)	–0.046 (1.125)	
Median house age (10 yrs.)	0.058 (0.403)	0.026 (0.175)	
Central city age (10 yrs.)	–0.006 (0.135)	0.012 (0.256)	
Constant	–1.045 (1.390)	–0.627 (0.900)	0.124 (0.419)
Log-likelihood value	–119.217	–118.177	–120.395
R^2	0.811	0.816	0.806

Notes: Each regression has 62 observations. Absolute *z*-values are in parentheses below the estimated coefficients. An asterisk indicates significance at the 5 percent level.

Source: Author's calculations based upon data from the U.S. Department of Transportation, Bureau of Transportation Statistics, *Census Transportation Planning Package*, and from the U.S. Department of Commerce, Bureau of the Census.

I include the average tract size in the regressions, because McMillen and Smith (2004) find that the subcenter identification procedure tends to find more subcenters when tract sizes are small. I include the last two variables, median house age and the age of the central city, because analysts such as Garreau (1991) have suggested that subcenters will come to dominate American cities in the future. Thus, newer cities may be more likely to have already developed subcenters. Median house age, as reported by the 1990 U.S. Census for 1990, is one measure of a city's age. I also use a variable suggested by Brueckner (1986) to measure city age: the number of years since the central city first reached 25 percent of its 1990 population level.

Table 5 displays the Poisson regression results. The estimated coefficients are interpreted as semi-elasticities. For example, the estimated coefficient for population in model 1 indicates that an additional million in population raises the expected number of subcenters by 14.8 percent. This estimate is stable across the three alternative model specifications, rising to 15.1 percent when I use *DVMTLANE* in place of the travel rate index to measure commuting cost and to 17.3 percent when I use only population and *DVMTLANE* as explanatory variables. The travel rate index and *DVMTLANE* have the expected positive signs, indicating that higher commuting cost leads to more subcenters. The coefficients for *DVMTLANE* indicate that an additional thousand miles traveled on average per mile of freeway lane raises the expected number of subcenters by 9.4 percent in model 2 and 9.3 percent in model 3.

The remaining explanatory variables are not important determinants of the number of subcenters in this sample. Metropolitan areas with large central cities tend to have fewer subcenters, but estimated coefficients for other explanatory variables—median income, average tract size, median house age, and age of the central city—are statistically insignificant. The pseudo-R^2s for the regressions (Cameron and Windmeijer, 1996) imply that the explanatory variables account for approximately 80 percent of the variation in the natural logarithm of the number of subcenters. Table 5 suggests a strong, simple empirical regularity in the number of subcenters in large metropolitan areas: The number of subcenters rises with population and commuting costs.

Conclusion

The traditional central business district is still the largest single employment site in most metropolitan areas. However, urban areas have become increasingly decentralized over time, and many cities now have more jobs in the suburbs than in the central city. Jobs are not spread randomly about the suburban landscape. Firms tend to locate at sites with ready access to the transportation system. Large employment subcenters have developed in many metropolitan areas that offer agglomeration economies to firms, while potentially reducing commuting times for suburban workers.

This article has documented the growth of employment subcenters in the Chicago metropolitan area between 1970 and 2000 and used forecasts of future employment to predict subcenter sites in 2020. A cluster analysis suggests that the employment mix in the subcenters has changed from predominantly manufacturing in 1970 to a mix that now closely resembles that of the overall metropolitan area. A regression analysis of employment density in the Chicago metropolitan area suggests that density rises near highway interchanges, rail stations, and along freight rail lines. Employment density also rises significantly in the area around employment subcenters.

Subcenters are found throughout the United States. Chicago had only 15 subcenters in 1990, New York had 38, and Los Angeles had 46. Of 62 large metropolitan areas analyzed in this article, 48 had at least one subcenter. The number of subcenters has a remarkably predictable pattern across the 62 urban areas. Poisson regression results imply that the number of subcenters rises with population and commuting costs. Thus, as cities grow, one can expect that subcenters will develop as firms congregate near intersections of major highways and in formerly satellite cities. Although new subcenters do not offer the same level of agglomeration economies as the traditional central city, they do offer lower land costs, easy access to highways, and the possibility of reduced wages for suburban workers whose commuting costs are reduced.

NOTES

[1]The tracts analyzed by Giuliano and Small (1991) are transportation analysis zones, as defined by the Southern California Association of Governments. The average area of the tracts is about 1.75 square miles.

[2]Performed using the program STATA.

[3]In 2000, 33.7 percent of the jobs in the Chicago metropolitan area were in the service sector, 19.3 percent were in retail, 11.3 percent were in manufacturing, 11.1 percent were in TCU, and 4.3 percent were in the government sector.

REFERENCES

Anas, A., and I. Kim, 1996, "General equilibrium models of polycentric urban land use with endogenous congestion and job agglomeration," *Journal of Urban Economics*, Vol. 40, No. 2, pp. 232–256.

Anderson, N., and W. Bogart, 2001, "The structure of sprawl: Identifying and characterizing employment centers in polycentric metropolitan areas," *American Journal of Economics and Sociology*, Vol. 60, No. 1, pp. 147–169.

Baumont, C., C. Ertur, and J. LeGallo, 2002, "Intra-urban spatial distributions of population and employment: The case of the agglomeration of Dijon, 1999," Université de Bourgone, manuscript.

Berliant, M., and H. Konishi, 2000, "The endogenous foundations of a city: Population agglomeration and marketplaces in a location-specific production economy," *Regional Science and Urban Economics*, Vol. 30, No. 3, pp. 289–324.

Bogart, W., and W. Ferry, 1999, "Employment centres in Greater Cleveland: Evidence of evolution of a formerly monocentric city," *Urban Studies*, Vol. 36, No. 12, pp. 2099–2110.

Booth, D., 1999, "Spatial patterns in the economic development of the Mountain West," *Growth and Change*, Vol. 30, No. 3, pp. 384–405.

Brueckner, J. K., 1986, "A switching regression analysis of urban population densities," *Journal of Urban Economics*, Vol. 19, No. 2, pp. 174–189.

Cameron, A. C., and P. K. Trivedi, 2001, "Essentials of count data regression," in *A Companion Guide to Theoretical Econometrics*, B. H. Baltagi (ed.), Malden, MA: Blackwell Publishers.

Cameron, A. C., and F. A. G. Windmeijer, 1996, "R-squared measures for count data regression models with applications to health-care utilization," *Journal of Business and Economic Statistics*, Vol. 14, No. 2, pp. 209–220.

Cervero, R. and K. Wu, 1998, "Subcentering and commuting: Evidence from the San Francisco Bay area, 1980–90," *Urban Studies*, Vol. 35, No. 7, pp. 1059–1076.

_____, 1997, "Polycentrism, commuting, and residential location in the San Francisco Bay area," *Environment and Planning A*, Vol. 29, No. 5, pp. 865–886.

Clark, C., 1951, "Urban population densities," *Journal of the Royal Statistical Society: Series A*, Vol. 114, pp. 490–496.

Combes, P., 2000, "Economic structure and local growth: France, 1984–1993," *Journal of Urban Economics*, Vol. 47, No. 3, pp. 329–355.

Craig, S. G., and P. Ng, 2001, "Using quantile smoothing splines to identify employment subcenters in a multicentric urban area," *Journal of Urban Economics*, Vol. 49, No. 1, pp. 100–120.

Fujita, M., P. Krugman, and T. Mori, 1999, "On the evolution of hierarchical urban systems," *European Economic Review*, Vol. 43, No. 2, pp. 209–251.

Fujita, M., and H. Ogawa, 1982, "Multiple equilibria and structural transition of non-monocentric urban configurations," *Regional Science and Urban Economics*, Vol. 12, No. 2, pp. 161–196.

Fujita, M., J. F. Thisse, and Y. Zenou, 1997, "On the endogenous formation of secondary employment centers in a city," *Journal of Urban Economics*, Vol. 41, No. 3, pp. 337–358.

Garreau, J., 1998, *Edge City: Life on the New Frontier*, New York: Doubleday.

Giuliano, G., and K. A. Small, 1991, "Subcenters in the Los Angeles region," *Regional Science and Urban Economics*, Vol. 21, No. 2, pp. 163–182.

Helsley, R., and A. Sullivan, 1991, "Urban subcenter formation," *Regional Science and Urban Economics*, Vol. 21, No. 2, pp. 255–275.

Henderson, J. V., and A. Mitra, 1996, "The new urban landscape: Developers and edge cities," *Regional Science and Urban Economics*, Vol. 26, No. 6, pp. 613–643.

Konishi, H., 2000, "Formation of hub cities: Transportation cost advantage and population agglomeration," *Journal of Urban Economics*, Vol. 48, No. 1, pp. 1–28.

McDonald, J. F., 1989, "Econometric studies of urban population density: A survey," *Journal of Urban Economics*, Vol. 26, No. 3, pp. 361–385.

_____, 1987, "The identification of urban employment subcenters," *Journal of Urban Economics*, Vol. 21, No. 2, pp. 242–258.

_____, 1985, "The intensity of land use in urban employment sectors: Chicago 1956–1970," *Journal of Urban Economics* Vol. 18, No. 3, pp. 261–277.

McDonald, J. F., and P. J. Prather, 1994, "Suburban employment centers: The case of Chicago," *Urban Studies*, Vol. 31, No. 2, pp. 201–218.

McMillen, D. P., 2001, "Nonparametric employment subcenter identification," *Journal of Urban Economics*, Vol. 50, No. 3, pp. 448–473.

McMillen, D. P., and J. F. McDonald, 2000, "Employment subcenters and subsequent real estate development in suburban Chicago," *Journal of Urban Economics*, Vol. 48, No. 1, pp. 135–157.

_____, 1997, "A nonparametric analysis of employment density in a polycentric city," *Journal of Regional Science*, Vol. 37, No. 4, pp. 591–612.

McMillen, D. P., and S. Smith, 2004, "The number of subcenters in large urban areas," *Journal of Urban Economics*, forthcoming.

Muñiz, I., A. Galindo, and M. García, 2003, "Cubic spline population density functions and subcenter delimitation: The case of Barcelona," *Urban Studies*, forthcoming.

Small, K. A., and S. Song, 1994, "Population and employment densities: Structure and change," *Journal of Urban Economics*, Vol. 36, No. 3, pp. 292–313.

Wieand, K., 1987, "An extension of the monocentric urban spatial equilibrium model to a multicenter setting: The case of the two-center city," *Journal of Urban Economics*, Vol. 21, No. 3, pp. 259–271.

Yinger, J., 1992, "City and suburb: Urban models with more than one employment center," *Journal of Urban Economics*, Vol. 31, No. 2, pp. 181–205.

Temporary help services and the volatility of industry output

Yukako Ono and Alexei Zelenev

Introduction and summary

Many firms today are changing their organizational structures by adopting more flexible staffing arrangements. Such arrangements frequently include hiring temporary workers, on-call staff, and private contractors. Recent surveys reveal that the use of flexible staffing arrangements and, in particular, the use of temporary workers in the U.S. economy has become more widespread.

According to a 1996 Upjohn survey of private employers, as many as 78 percent of establishments used at least one type of flexible staffing arrangements in 1996; 46 percent of establishments employed temporary workers (Houseman, 2001a). The Bureau of Labor Statistics (BLS) data on employment reveal that the temporary help service (THS) industry, which supplies temporary workers, grew by more than 700 percent between 1982 and 2000—THS employment increased from approximately 417,000 to 3,489,600 in that period. The dramatic increase in the use of temporary workers has generated a vigorous debate among economists and policymakers about the costs and benefits of flexible staffing arrangements.

One of the most frequently cited reasons for the adoption of flexible staffing arrangements is that such arrangements allow firms to accommodate unexpected increases and decreases in business activity. By using flexible labor, firms, especially in volatile industries, can meet a surge in demand more efficiently; and if business activity experiences a downturn, firms can reduce their flexible work force without making costly adjustments to their permanent employment levels.

However, very few studies offer direct empirical evidence to support this view. The relationship between the rise and fall in a firm's output and its use of flexible staffing arrangements is not as straightforward as it might seem at first. On one hand, the volatility of output may induce firms to expand the use of flexible staffing arrangements, increasing the aggregate number of flexible workers. But on the other hand, if the demand for flexible labor fluctuates a great deal in response to firms' hiring and laying-off patterns, subcontractors and agencies supplying temporary help might find it difficult to continue providing such services to the market, potentially decreasing the use of flexible labor.

In this article, we conduct a closer examination of the relationship between the fluctuations of output and labor supplied by THS agencies, one of the commonly used forms of flexible staffing arrangements. Using state-level data, we analyze the shares of THS employment in relation to the output volatility of other sectors (non-THS industries) across the U.S. from 1977 to 1997. In order to capture the effect of volatility, we construct an index that measures the degree of fluctuation of industry output in each state. Furthermore, we decompose the volatility index into two components: one that measures the volatility associated with each individual industry; and a second component that measures the *co-movement* of output fluctuations for different industries in the same state. We find evidence that there is a positive association between the level of output volatility and the share of temporary service employment across different states. This result suggests that industries that experience greater fluctuations in output use more THS labor than industries that are relatively stable. Furthermore, we find that the THS shares are lower in the states in which the fluctuations of output are highly correlated among industries, suggesting

Yukako Ono is an economist and Alexei Zelenev is an associate economist at the Federal Reserve Bank of Chicago. *The authors are grateful to their colleagues at the Federal Reserve Bank of Chicago for helpful comments. They particularly thank Dan Aaronson, Helen Koshy, David Marshall, and Dan Sullivan for many detailed comments and suggestions on earlier drafts.*

that the flexibility of labor markets is lower in areas with a high degree of co-movement of output fluctuations across different industries. One possible interpretation is that THS agencies may find it difficult and costly to supply temporary workers to the labor market in areas with a high degree of industry co-movement, where many client firms simultaneously reduce or increase their usage of temporary workers.

THS industry: Trends and issues

The THS industry has been one of the fastest growing industries in the U.S. economy, outpacing many traditional industry sectors (Clinton, 1997). Analysis of recent data surveys reveals that almost all sectors of the economy have expanded their usage of temporary workers. Based on estimation in Estẽvao and Lach (1999), the biggest increases have been in the manufacturing and service sectors; in particular, by 1997, close to 4 percent of employees in manufacturing were employees of THS firms. Other sectors, notably finance, insurance, and trade and construction, have experienced substantial gains over time as well. Although temporary positions often involve clerical and administrative work (more than one-third of all temporary workers hold administrative/clerical positions [Cohany, 1998]), temporary workers represent a wide range of occupations. From lawyers to physicians, from manufacturing to construction workers, the THS industry supplies temporary workers with a diverse range of skills and work experience to the labor market (Rogers, 2000).

THS agencies come in a variety of sizes. Among the largest are Adecco SA, Kelly Services Inc., and Manpower Inc, each of which operates between 2,500 and 5,500 offices in the U.S. and around the world. But there are also many smaller agencies. According to the *1992 Enterprise Statistics* report (U.S. Department of Commerce, Bureau of the Census, 1992), there were 22,223 companies with a total of 32,515 offices that engaged in some kind of personnel supply services in the U.S. Some companies are highly specialized and provide highly skilled workers, such as biological scientists and engineers, while other companies provide workers with more general skills, such as administrative assistants and other office staff (Rogers, 2000).

THS agencies can enhance the efficiency and flexibility of the labor markets in a number of ways. The presence of THS agencies in a region reduces job-search costs and informational asymmetries by helping to match the workers who are looking for a temporary work opportunity with the firms that need temporary help.

For many people, THS employment presents a direct alternative to labor force withdrawal or unemployment. Working for a THS agency may also grant workers more freedom of choice by allowing them to combine work with other activities, such as child-rearing or study, for example. For others, temporary work opportunities can become a route to full-time employment; this route may be especially appealing for workers with little previous experience and/or training. According to a recent study, more than half of employees in temporary positions find permanent jobs within one year of their first interview (Segal and Sullivan, 1997). In addition, THS agencies screen and train their workers. The resulting skills and knowledge may increase workers' productivity and signal to client firms that the workers are motivated and fully qualified; this, in turn, may lead to opportunities for full-time employment. Firms' increased use of THS also implies that the demand for worker screening may be rising (Autor, 2001).

In addition to the advantages of THS employment outlined above, however, there are a number of costs and limitations. On average, temporary workers in nonprofessional categories receive much lower wages than permanent workers, although they frequently perform the same tasks as permanent staff members (Segal and Sullivan, 1995 and 1998). In addition, some temporary workers work on a permanent basis (so-called "perma temps") without receiving the same benefits and wages as permanent workers. The law does not offer the same protection to temporary workers as it does to permanent employees.[1]

For the client firms, the use of THS provides a number of benefits. The public and private sectors gain the advantage of drawing fairly easily and quickly on temporary workers when confronted by unexpected departures and absences among their permanent work force. The use of temporary workers also allows firms to accommodate fluctuations in business activity more efficiently, for example, a sudden increase or drop in product demand. In fact, more than half the establishments surveyed in 1996 by the Upjohn Institute listed "[the ability of THS agencies to] provide needed assistance at the time of unexpected increase in business [activity]" among their top reasons for using temporary workers (Houseman, 2001a). Increasing costs associated with laying off permanent workers might have led firms to seek flexibility by hiring temporary workers. According to Autor (2003), between 1973 and 1995, 46 states adopted exceptions to the common law doctrine of employment, which limited employers' discretion to fire permanent workers and made them vulnerable to potentially costly litigation; Autor's study found that this change to the legal environment explained 20 percent of the growth of THS during this period.

Flexible labor and volatility: Some evidence

Many economists have noted that the demand for THS employment is very sensitive to the business cycle. Segal and Sullivan (1997) interpret the cyclical sensitivity of the THS industry as an indicator that it provides a buffer for firms that face high costs of adjusting permanent employment. They argue that the flexibility granted by the use of THS workers, coupled with firms' reluctance to adjust their levels of permanent employment, is one of the reasons THS employment is much more volatile than aggregate employment, falling more during contractions and rising more during expansions.

While much research has pointed out the importance of THS labor in helping firms to accommodate fluctuations in output demand more efficiently, there have been few empirical studies that looked at the extent to which temporary labor facilitates flexibility or that have analyzed the association between output volatility and the use of temporary labor. Within the research that does exist, the evidence on the question of whether more volatile industries use more flexible staffing arrangements has been rather mixed and, in some instances, inconsistent.

An example of a study that looks at the relationship between fluctuations of output and THS employment is Golden (1996). Golden finds evidence that a rise in demand for output above the long-run trend produces a strong concurrent rise in demand for temporary labor. Her work suggests that temporary employment facilitates flexibility and allows firms to meet short-term fluctuations in demand and avoid costly adjustments to permanent employment. The evidence she presents seems to be consistent with the buffering hypothesis we mentioned earlier.

However, other studies suggest that greater volatility of output does not increase and might actually decrease firms' demand for flexible staffing arrangements. One example is a paper by Abraham and Taylor (1996), in which they analyze manufacturing establishments' practices of outsourcing business services. Although their analysis does not focus on temporary agencies directly, their study has important implications for understanding the use of THS services, since hiring THS workers can be considered as a kind of outsourcing activity. Using the seasonal fluctuations of industry employment as a proxy for volatility of demand, Abraham and Taylor (1996) find that establishments in more volatile industries appear less likely to contract out various services. In particular, they find that the probability of outsourcing janitorial, machine maintenance, engineering and drafting, and accounting services, on average, decreases as the degree of seasonal volatility rises. Their findings run counter to the story that firms use subcontractors and temporary workers in order to smooth the flow of in-house work during peak periods.

In sum, the existing literature provides mixed evidence for the association between output volatility and the use of flexible staffing arrangements. In reality, whether volatility of product demand will increase or decrease a firm's use of temporary workers may depend on many factors. While greater volatility of output might create greater demand for THS workers, if the demand for temporary workers fluctuates a lot, THS agencies might find it difficult and costly to supply temporary workers to the labor market. For example, during a downturn, THS agencies might face a risk of not being able to reallocate temporary workers from one industry to another, if many client firms are simultaneously reducing their usage of temporary employment; during periods of expansion, THS agencies may have to put more effort into finding suitable matches of temporary workers and clients. As a result, THS agencies may charge a higher premium to client firms, which may make the option of hiring temporary workers less cost-effective.

In this article, we investigate whether there is any evidence for the two different roles that volatility plays in determining the degree of THS usage, by examining the cross-sectional relationship between THS employment share and other sectors' output volatility across U.S. states. In particular, we examine whether there is any evidence that the use of THS is offset by correlated patterns of output fluctuations among industries. And to do this, we calculate a volatility index. In the next section, we describe the procedure that we use.

Measuring output volatility at state level

The amount of goods that firms produce varies from year to year. Firms adjust their production levels in response to changes in market conditions. Changes in consumer demand, as well as changes in the costs of production, can generate positive or negative shocks, resulting in either growth or contraction of industry output. Shocks can be industry-specific, affecting the level of output in one particular industry, or shocks can be common to more than one industry, affecting the level of output of several industries, sometimes in different sectors of the economy. Examples of industry-specific shocks include technological innovation and changes in the price of inputs, which affect industry production; examples of common shocks include changes in interest rates and taxes, which affect the ability of firms in many sectors to borrow and invest in infrastructure.

Fluctuations in output across many industries often are the result of a common shock. The resulting co-movement of output fluctuations of industries that make up a state's economy would comprise an important part of the state's overall output fluctuation. For example, output fluctuations in textile industries are more highly correlated with fluctuations in the apparel industry than in the printing industry. So, *ceterus paribus*, a state with high shares of apparel and textiles is more volatile, on average, than a state that has equally large shares of apparel and printing. The co-location of negatively correlated (or even uncorrelated) industries in a state can produce a kind of stabilizing effect, potentially lowering the volatility of demand for THS and providing a better environment for THS agencies to operate. To capture such an effect, using a method from Conroy (1975) and Diamond and Simon (1990), we decompose the volatility of output into two parts: one part that results from each industry's output fluctuation and another part that results from the correlation of output fluctuations.

To compute the volatility index, we use industry output rather than industry employment, because the size of the permanent work force in each industry can be directly related to the number of temporary workers each industry decides to use. If firms in volatile industries use temporary workers to reduce fluctuations in the permanent staff, then using employment to measure volatility would not uncover any volatility, because permanent employment would not change. (Temporary workers supplied by THS agencies are on the payroll of the THS industry and are not included in employment of the client industry in our data.) Below, we describe the construction of the index in more detail. First, we show how we capture the volatility of each industry's output; then, we show how we compute the volatility for each state in each year.

First, we decompose the growth rate of each industry's output into two components: a secular component g_{it}, which captures the trend growth path, and a cyclical component $\tilde{g}_{it}$, which deviates from the trend value. We call the latter the residual growth rate. The growth rate of industry i in year t can be written as,

1) $\textit{growth rate}_{it} = g_{it} + \tilde{g}_{it}$.

Figure 1 provides an illustration of the relationship between the growth rate and the residual growth rate.

The rise and fall of the residual growth rate $\tilde{g}_{it}$ over time captures the fluctuation of output for industry i, so we use the residual growth rates in our calculations of the index. To obtain the residual growth rate, we use the Bureau of Economic Analysis' (BEA) real gross domestic product (GDP) data for 54 industries[2] for the period between 1978 and 2001. We regress the real growth rate on time for each industry and retrieve the residual terms by taking the difference between the predicted and actual growth rate values.[3] Note that in many industries, the real growth rate of output fluctuates around a certain constant value. In such cases, the residual growth rate will be almost the same as the deviation from the average real growth rate. However, for some industries, there are steady upward or downward changes in the real growth rate during this period, which are captured by the coefficient of time variable. For example, in the case of the food product industry, while the real growth rate moved up and down, on average it was declining between 1978 and 2001. The coefficient of time variable was –0.00382 and statistically significant at the 5 percent level.

FIGURE 1

Relationship between trend growth rate and residual growth rate

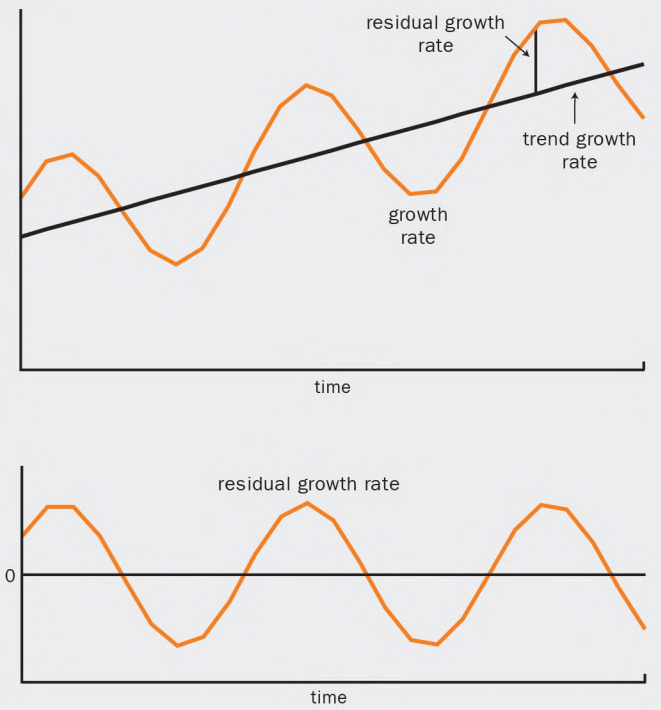


Figure 2 shows the residual growth rates for several selected industries in the manufacturing and service sectors. It is well known that manufacturing industries are more volatile than services. In figure 2, the residual growth rates of manufacturing industries move over much greater ranges than those of service industries.

Next, we use the residual growth rates for each industry to calculate the overall growth of the state economy, which we need to compute our measure of volatility. Since each state has an assortment of many industries, to capture the residual growth rate at the state level, we take a weighted average of the residual growth rates of each industry, treating the industry's employment share in each state as weights. So the state-level residual growth rate, $\tilde{g}_{st}$, can be written as:

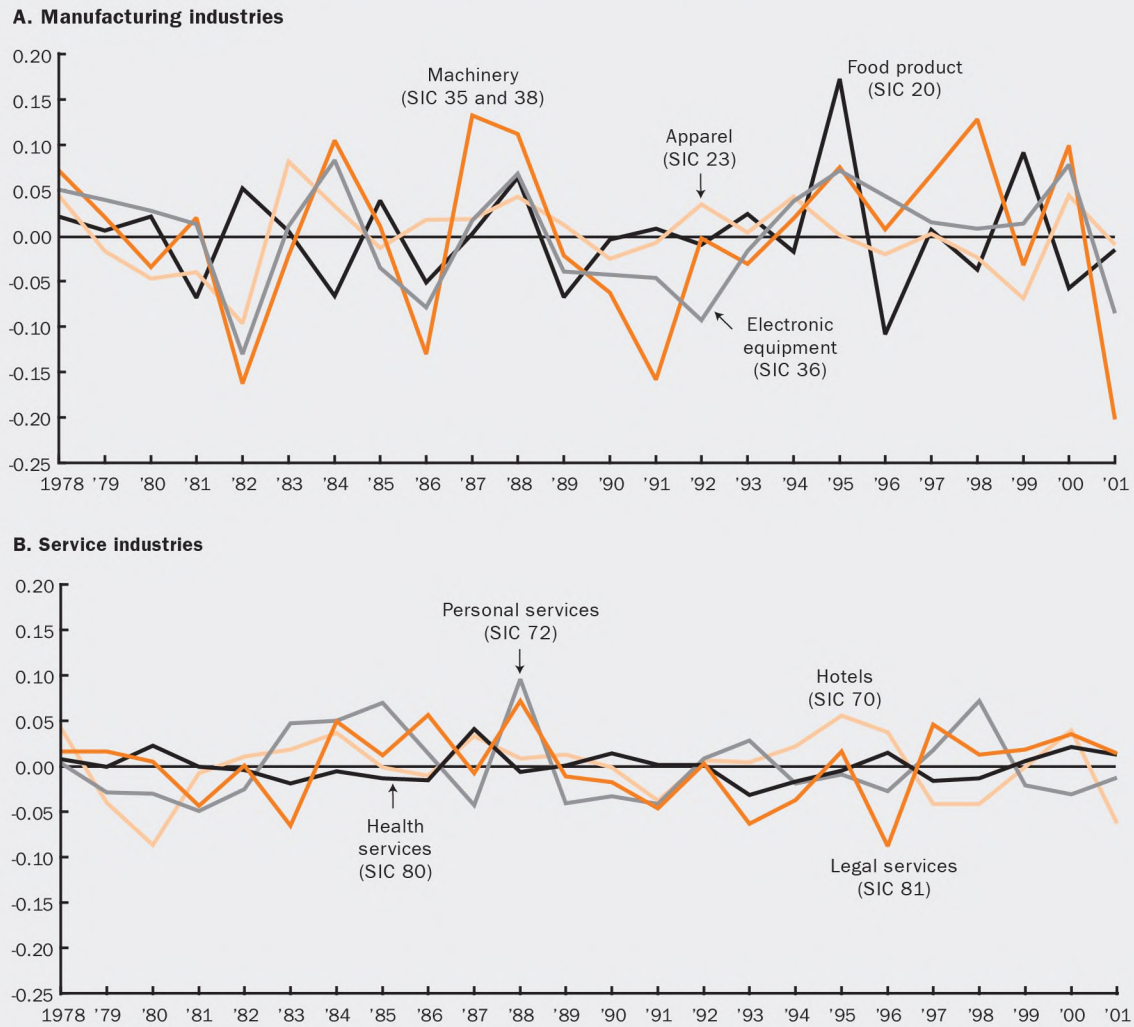
$$2) \quad \tilde{g}_{st} = \sum_i S_{ist} \tilde{g}_{it},$$

where S_{ist} is industry i's share in non-THS employment in state s in year t.[4]

One measure of fluctuations frequently used by economists is variance, which captures the dispersion in the data. Thus, we calculate the variance of the weighted averaged residual growth rates (for each

FIGURE 2

Residual growth rates of output for selected industries



Source: Authors' calculations based on data from the U.S. Department of Commerce, Bureau of Economic Analysis.

state) to quantify the level of output volatility in each state. If the industry fluctuations are independent, the variance of averaged growth rates at the state level, VAR_{st}, can be written as:

$$3)\quad VAR_{st} = \sum_i S_{ist}^2 \sigma_i^2 ,$$

where σ_i^2 is the variance of the residual growth rate of industry i. However, output fluctuations in many industries are actually correlated. In such a case, VAR_{st} will have an additional component, and is written as:

$$4)\quad \underbrace{VAR_{st}}_{\text{Overall volatility}} = \underbrace{\sum_i S_{ist}^2 \sigma_i^2}_{\text{UVAR}} + \underbrace{\sum_i \sum_{i \neq j} S_{ist} S_{jst} \sigma_{ij}}_{\text{CVAR}} ,$$

where σ_{ij} represents the covariance of the residual growth rates of industries i and j. We refer to $\sum_i S_{ist}^2 \sigma_i^2$ as the uncorrelated output variance (*UVAR*) and $\sum_i \sum_{i \neq j} S_{ist} S_{jst} \sigma_{ij}$ as the co-movement variance (*CVAR*). *UVAR* measures the volatility when output fluctuations are not correlated among industries, while *CVAR* measures an additional component to the volatility that results when output fluctuations are correlated among industries. In the actual computation of the index, we use sample variances and sample covariance of residual growth rates, which we calculate based on each industry's residual growth rates from 1978 to 2001.

While the output of many industries tends to move together, the degree to which the output fluctuations coincide differs across industries. For example, figure 3, panels A, B, and C show the residual growth rates of each industry for 1982, 1983, and 1984, respectively. During this period, the U.S. economy experienced both recession and expansion. Based on the aggregate GDP data from the BEA, between 1981 and 1982, the real GDP growth rate was –2.02 percent, which was followed by some recovery in 1983 and further expansion in 1984, resulting in real GDP growth of 7.26 percent between 1983 and 1984. Such changes in the growth of the overall economy are reflected in the residual growth rates that we calculated. As shown in figure 3, panels A and C, for most industries, the residual growth rates are negative in 1982 and positive in 1984. However, some industries' growth paths were moving in the opposite direction from most of their peers. Moreover, figure 3, panels A and B show that between 1982 and 1983, the growth rate changed from negative to positive for some industries and remained negative for others. If a state has a majority of industries whose output fluctuates together, this will increase the state's overall volatility. In contrast, if a state has mostly industries whose output fluctuations do not coincide, this will stabilize the overall volatility. Box 1 provides an illustrative example of how the correlation of output fluctuations among industries influences a state's overall output volatility.

Table 1 shows the five most volatile and five least volatile states based on our volatility measures. Between 1977 and 1997, it appears that, on average, output volatility is the highest in Indiana, with an overall output variance of .000760. That is, Indiana's output growth rate deviates from its trend by 2.76 percent, on average. In contrast, the District of Columbia appears to have the lowest volatility at .000231 on average, translating to 1.52 percent deviation of its growth rate from the trend. Table 1 also shows that Indiana experienced the highest volatility, and most of that volatility (81 percent) resulted from the co-movement of output fluctuations among the state's industries. To see how the composition of overall volatility can vary across different states, we can compare North Dakota and Ohio. Overall volatility in Ohio is much greater than in North Dakota. However, table 1 shows that the greater volatility in Ohio relative to North Dakota is due to the greater *CVAR* in Ohio (*UVAR* is almost the same for the two states)—in other words, Ohio is more "volatile" than North Dakota because of the higher degree of co-movement exhibited by Ohio's industry mix.

Empirical specification and data

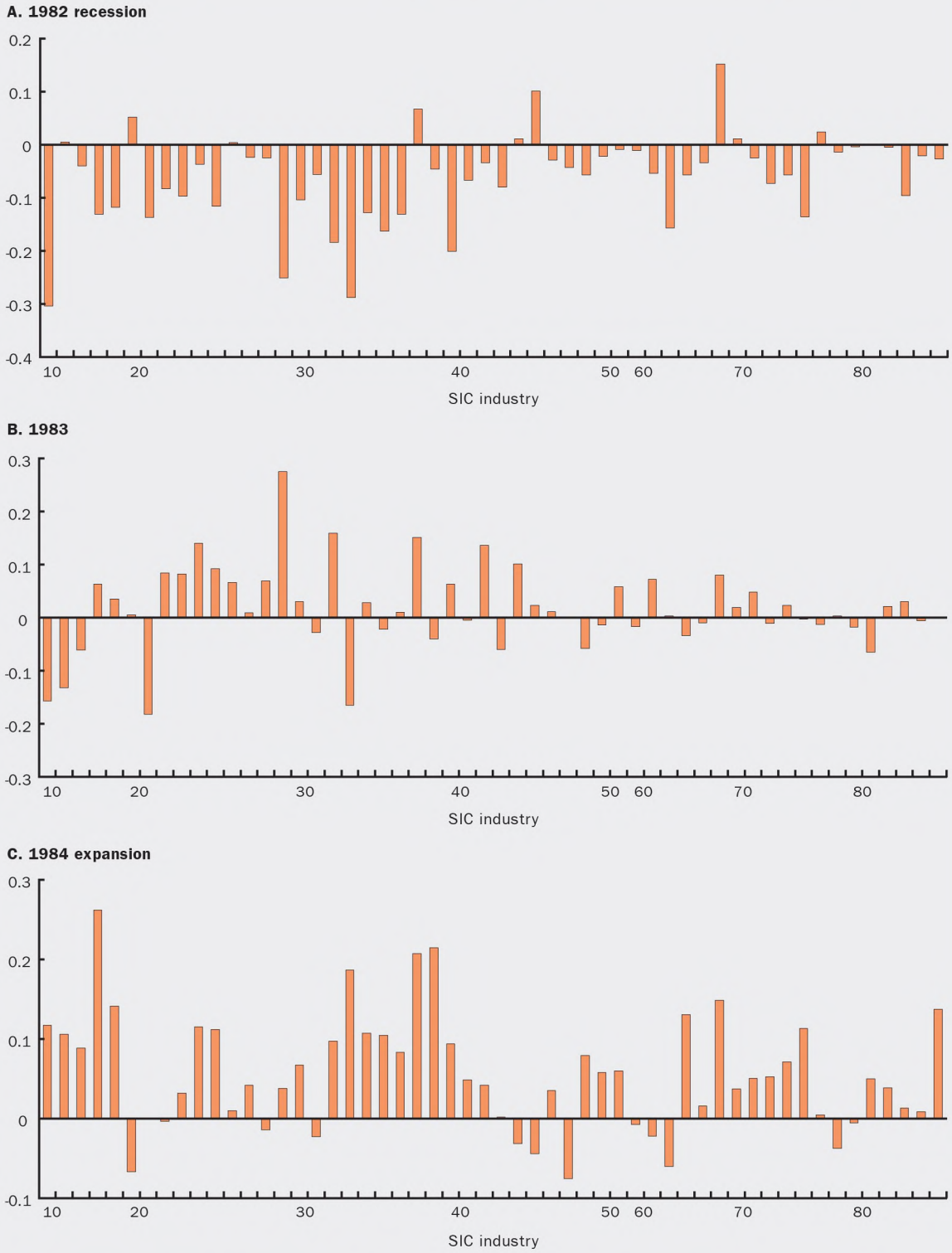
Using the volatility measure calculated above, we examine how each component of the volatility measure is associated with the share of THS employment in each state. We proceed with the following specification:

$$5)\quad THSshare_{st} = (UVAR_{st}, CVAR_{st}, \ln non\text{-}THS\ emp_{st}, Urate_{st}, X_{st}, \text{Year dummies})\ \beta + u_{st},$$

where $THSshare_{st}$ represents the THS employment share in state s in year t, β is a vector of coefficients, and u_{st} is a random component. In the regression, we control for the size of the labor market in each state by including the size of non-THS employment (*non-THS emp*) in logarithm. In a larger labor market, each THS agency may have a longer list of workers seeking temporary work. This might facilitate scale economies for THS agencies in their searching process and allow them to provide their services more efficiently; this in turn might increase the use of THS. We also control for the state's unemployment rate (*Urate*); a higher unemployment rate might reduce employment opportunities for temporary workers more than for permanent workers and might in turn influence the THS employment share.

FIGURE 3

Residual growth rates of output by industry for years of recession and expansion



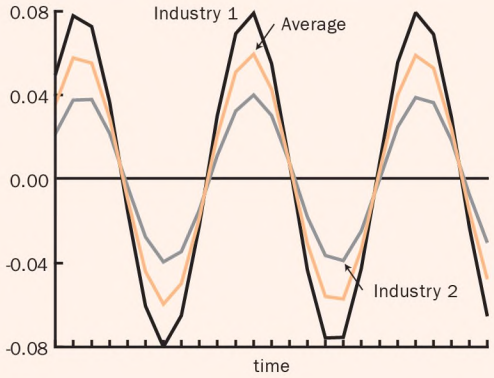
Source: Authors' calculations based on data from the U.S. Department of Commerce, Bureau of Economic Analysis.

BOX 1

An example of how the co-movement of output fluctuations affects the overall volatility

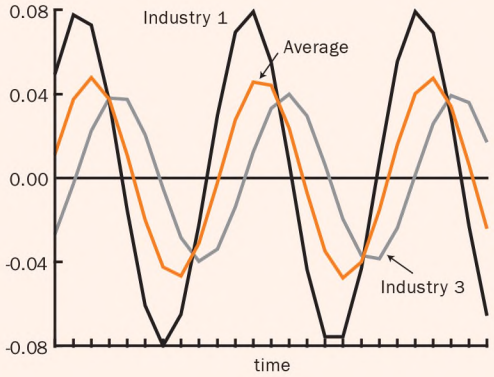
We consider the following hypothetical case for illustrative purposes. Take two states, A and B. State A's economy consists of two industries, industries 1 and 2, and they are about equal in size in state A. We assume that industry 1 is more volatile than industry 2 (the range over which the output of industry 1 fluctuates is wider), but that their output moves up and down together (co-movement). So these two industries expand and contract at approximately the same time. Figure B1 illustrates the co-movement of the (detrended) growth rates of output in industries 1 and 2. Industry 1's growth rate typically fluctuates from its trend growth rate by –8 percent to 8 percent; and industry 2's, by –4 percent to 4 percent. Since, in state A, the two industries are equal in size, the average residual growth rate in state A is simply the mean of the residual growth rates of industries 1 and 2.

Figure B1
Residual growth rate of output, state A



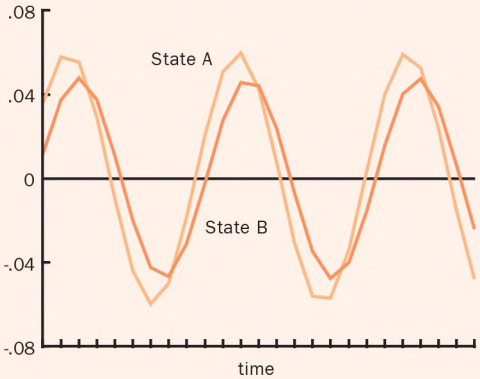
In state B, we assume that half of the economy is represented by industry 1 as in state A and the other half by another industry, industry 3. As we show in figure B2, industry 3's output is as volatile as that of industry 2—industry 3's growth rate deviates from its trend by almost the same degree as that of industry 2. However, unlike industry 2, industry 3's output fluctuation does not coincide with that of industry 1. In state A, the correlation of residual growth rates between industries 1 and 2 is very close to 1. However, in state B, the correlation between residual growth

Figure B2
Residual growth rate of output, state B



rates of industries 1 and 3 is about 0.1. Because of the relatively low degree of co-movement of output fluctuations in industries 1 and 3, on average the output is less volatile in state B than in state A. In figure B3, we plot the average of the residual growth rates in states A and B. The average residual growth rate in state A ranges between –6 percent and 6 percent, and that in state B ranges between –4.5 percent and 4.5 percent. So the co-movement of output fluctuations between industries and industrial composition matter for the overall volatility in each state.

Figure B3
Comparison between states A and B



We control for factors that may influence the supply of THS workers by including the demographic characteristics of each state (share of population by age, sex, and race). We also include year dummies to control for the increase in THS share that every state has experienced. After controlling for these variables, we expect the coefficient of *UVAR* to be positive and that of *CVAR* to be negative, since as we discussed before, greater volatility would increase the demand for temporary workers, while greater correlation of output

TABLE 1

States with highest and lowest volatility

	VAR	UVAR	CVAR
Five states with highest volatility			
Indiana	.000760	.000144	.000616
Michigan	.000757	.000168	.000589
Connecticut	.000700	.000153	.000547
Ohio	.000695	.000136	.000559
New Hampshire	.000692	.000158	.000534
	VAR	**UVAR**	**CVAR**
Five states with lowest volatility			
Alaska	.000433	.000163	.000270
South Dakota	.000430	.000137	.000293
Hawaii	.000427	.000143	.000284
North Dakota	.000386	.000139	.000247
District of Columbia	.000231	.000093	.000138

Notes: VAR is overall volatility; UVAR is uncorrelated output variance; and CVAR is co-movement variance.
Source: Authors' calculations based on data from the U.S. Department of Commerce, Bureau of the Census, *County Business Patterns* and the U.S. Department of Commerce, Bureau of Economic Analysis.

fluctuations among industries may shift down the supply curve of temporary workers and lower the use of temporary workers.

The data on employment for the THS and non-THS sectors are taken from *County Business Patterns* (CBP) 1977–97.[5] The CBP reports are published by the U.S. Department of Commerce, Bureau of the Census and provide county- as well as state-level industry data, based on the four-digit Standard Industrial Classification (SIC) codes. We use state unemployment time-series data from the BEA. In addition, we use the Census population data and the *Current Population Survey* (CPS) for the demographic profiles of each state from 1977 to 1997. In particular, we calculate the shares of population in different age groups and the shares of female and black population in each state and each year and include them in our regression.

Table 2 shows the summary statistics of the dependent variable and covariates of 50 states and the District of Columbia. On average, THS employment made up 0.98 percent of state employment between 1977 and 1997.[6] Within the same period, each state, on average, had about 1.6 million people employed in the non-THS private sector and an unemployment rate of 6.4 percent. Between 1977 and 1997, the share of people under 17 years of age averaged about 27.2 percent in each state; the share of people aged 18 to 24, 11.5 percent; the share of people aged 25 to 64, 49.1 percent; and the share of those aged 65 and over, 11.9 percent.

According to the CBP data, there is a lot of variation in THS employment across different states. Table 3 shows the top five and bottom five states in terms of the average THS employment shares from 1977 to 1997. On average, in Florida, THS employment represented about 2 percent of total state employment, while in North Dakota it was only about 0.2 percent. Figure 4, panels A and B show cross-sectional variation in THS employment shares in 1977 and 1997. While the increase in the THS employment share is a nationwide phenomenon, the growth of THS employment seems to vary across the U.S. For example, Arkansas, Oregon, and Utah have some of

TABLE 2

Summary statistics

Variable	Mean	Standard deviation
THS employment share	.00983	.00792
Covariates		
Employment of non-THS sectors	1,601,183	1,775,804
Unemployment rate (share)	.0644	.0209
Volatility measure		
Overall volatility (VAR)	.000562	.000113
Uncorrelated output variance (UVAR)	.000136	.000025
Co-movement variance (CVAR)	.000426	.000105
Demographic variables: Share of state population		
Age under 17	.272	.028
Age 18 to 24	.115	.017
Age 25 to 64	.491	.028
Age 65 or more	.119	.022
Female	.512	.010
Black	.105	.124

Notes: THS is temporary help service; VAR is overall volatility; UVAR is uncorrelated output variance; and CVAR is co-movement variance.
Source: Authors' calculations based on data from Haver Analytics, from the U.S. Department of Commerce, Bureau of the Census, *County Business Patterns*, *Current Population Survey*, and population census, and from the U.S. Department of Commerce, Bureau of Economic Analysis.

the nation's fastest growing THS sectors, while other states such as New York and Washington have shown more modest rates of increase. (We also compared the THS employment shares between 1987 and 1988 and found that the relative levels of THS employment shares across the states are very similar between these two years, which suggests that the 1987 SIC change is not likely to be an important factor in producing the differences in panels A and B.) The comparison between panels A and B in figure 4 reveals that the regional composition of temporary employment might have shifted away from the North East toward the South West over the 20-year period we study. In the next section, we examine how the cross-sectional differences in THS employment shares are related to output volatility at state level.

Results

In this section, we discuss the results from our regression analysis.[7] In column 1 of table 4, we consider the effects of both *UVAR* and *CVAR* on the shares of temporary employment across the states. We find that the coefficient for *UVAR* is positive and that of *CVAR* is negative, which is consistent with our hypothesis outlined above. Both coefficients are significant. The empirical findings do not change qualitatively when we allow the effect of each demographic component to vary over time (regression in column 2).

The positive coefficient for *UVAR* suggests that there may be greater demand for temporary labor in states with a mix of volatile industries. It is possible that volatility of output among industries in these states creates more business opportunities for THS agencies, which might attract more agencies to the local market and enhance competition among them. As a result of greater competition, the price charged to client firms is likely to fall, which in turn may increase the use of THS.

However, the negative coefficient for *CVAR* indicates that, for a given level of *UVAR*, THS employment shares are lower if output fluctuations tend to coincide across industries. These results intuitively make sense. First, if output fluctuations are highly correlated among industries, decisions to hire and fire temporary workers are more likely to be correlated among industries as well. In such a case, the demand for temporary workers that each THS agency faces will become more volatile. As a result, THS agencies might find it more costly to provide a matching service in a timely manner. The increase in the costs of matching may be reflected in higher prices charged to client firms, making the use of temporary labor less attractive. Second, the co-movement of output fluctuations might also reduce the supply of temporary labor. If all industries decide to reduce their use of temporary workers simultaneously as a result of a common shock to production, THS agencies will find it difficult to place their workers. Thus, temporary workers might face a higher risk of not being able to secure an alternative assignment once the current assignment ends. This might make temporary work less attractive, leading to a lower supply of temporary labor and a lower quality of services offered by THS agencies. As a result, client firms might use THS services less intensively.

Finally, depending on the sample, in some cases the effect of *CVAR* may dominate the effect of *UVAR*, which may result in a negative correlation between overall volatility and THS employment share. In our sample, as shown in the regression in column 3 of table 4, on average, the positive effects of *UVAR* and the negative effects of *CVAR* seem to offset each other; the effects of overall volatility (*VAR*) on temporary service employment appear to be insignificant at the 10 percent level.

TABLE 3

THS employment share, average 1977–97

U.S.	.983 (%)
Top five states	
Florida	2.011
Arizona	1.582
California	1.563
Texas	1.511
District of Columbia	1.461
Bottom five states	
Alaska	.360
Montana	.325
Wyoming	.323
South Dakota	.321
North Dakota	.198

Note: THS is temporary help service.
Source: Authors' calculations based on data from the U.S. Department of Commerce, Bureau of the Census, *County Business Patterns*.

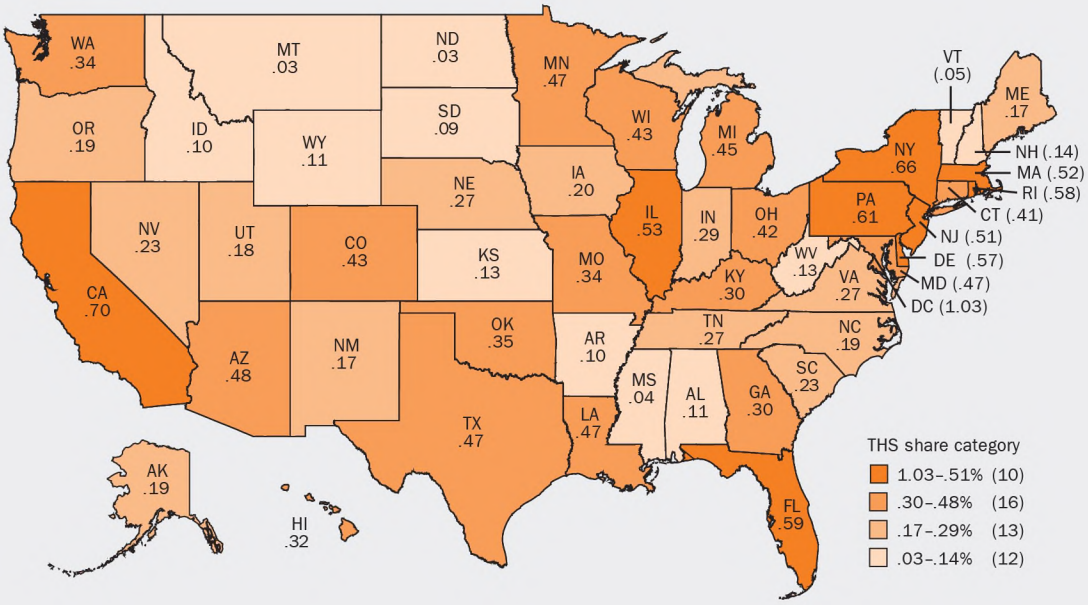
Effects of other variables

In addition to volatility, we examine the effects of unemployment and demographic variables on THS employment share across the U.S. The unemployment rate appears to be negatively related to THS employment share. This may be connected to the fact that temporary workers may be used as buffers—a decrease in the use of temporary workers during a downturn would contribute to a higher unemployment rate. The result is also consistent with Otto (1999), who finds that the share of temporary employment reduces the natural rate of unemployment in local labor markets.

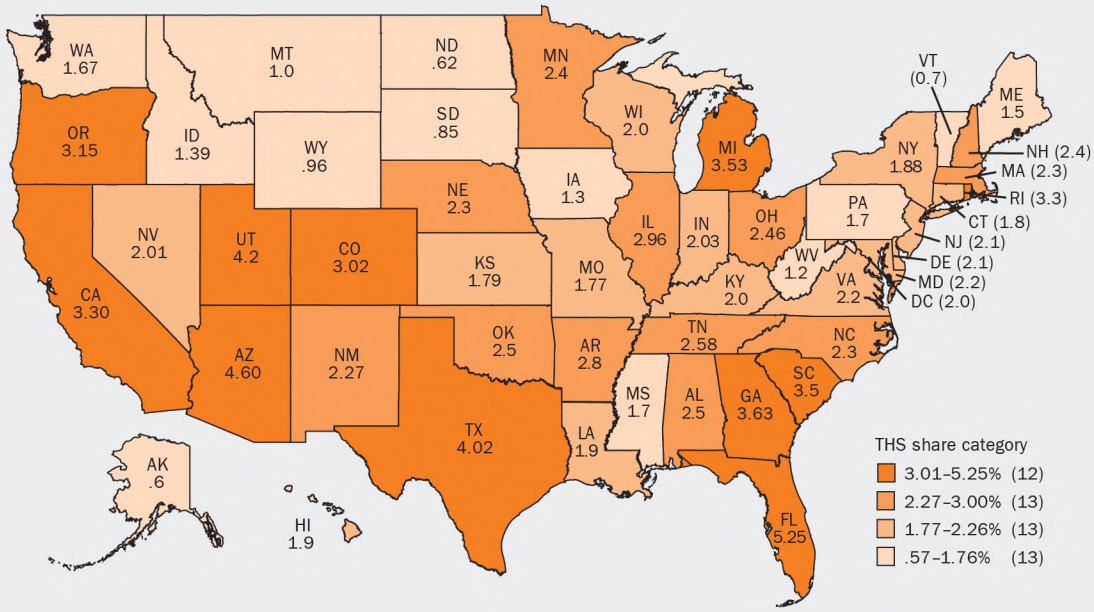
FIGURE 4

THS employment percentage of total state employment

A. 1977



B. 1997



Source: Authors' calculations based on data from the U.S. Department of Commerce, Bureau of the Census, *County Business Patterns*.

In addition, we find that state demographic characteristics appear to have an effect on the supply of temporary workers. In particular, large shares of THS employment are positively associated with higher shares of female, black, and 18–24 year old population groups. This result is consistent with other studies that analyze the demographic composition of the temporary work force (Polivka, 1996, and Cohany, 1998).

It is also interesting to see how the effects of demographic factors change over time between 1977 and 1997. In particular, in regression 2 in table 4, the coefficient for the interaction term between black population share and time (*Black* × *T*) turns out to be positive and significant, suggesting that more black workers were involved in temporary work in 1997 than in 1977. In addition, we find that the interaction term between the percentage of children (that is, share of population under age 17) and time obtains a positive and significant coefficient, suggesting that households with children were more likely to be involved in temporary labor in 1997 than in 1977.[8]

TABLE 4

Effect of volatility measure

	(Dependent variable: THS employment share)		
	1	2	3
UVAR	14.893***	10.772**	
	(4.980)	(2.110)	
CVAR	–2.194*	–3.081***	
	(1.218)	(1.190)	
VAR			–.146
			(1.059)
Control variables			
Log employment of all other sectors	.00191***	.00189***	.00171***
	(.000115)	(.000111)	(.000103)
Unemployment rate	–.0254***	–.0238***	–.0247***
	(.00554)	(.00541)	(.00556)
Demographic characteristics, share of state population			
Age 17 and under	–.0320***	–.0587***	–.0338***
	(.00484)	(.00951)	(.00483)
Age under 17 × T		.00397***	
		(.000887)	
Age 18 to 24	.0391***	.0740***	.0334**
	(.0135)	(.0253)	(.0134)
Age 18 to 24 × T		.000280	
		(.00203)	
Age 65+	–.0378***	–.0157	–.0389***
	(.00716)	(.0116)	(.00719)
Age 65+ × T		–.00423***	
		(.000913)	
Female	.0482***	.0548***	.0311**
	(.0168)	(.0171)	(.0159)
Female × T		.000125	
		(.0000853)	
Black	.00323**	–.00329*	.00317***
	(.00113)	(.00179)	(.00113)
Black × T		.000660***	
		(.000152)	

Notes: THS is temporary help services; VAR is overall volatility; UVAR is uncorrelated output variance; and CVAR is co-movement variance. Year dummies are included in the regression. T = (Year–1977). Standard errors are in parentheses. * Indicates significant at 10 percent level; ** indicates significant at 5 percent level; and *** indicates significant at 1 percent level.

Source: Authors' calculations based on data from Haver Analytics, from the U.S. Department of Commerce, Bureau of the Census, *County Business Patterns*, *Current Population Survey*, and population census, and from the U.S. Department of Commerce, Bureau of Economic Analysis.

Conclusion

Many researchers have argued that the presence of the THS industry enhances flexibility in labor markets by allowing firms to accommodate cyclical fluctuations in output demand more efficiently. In this article, we analyze the relationship between output volatility and the use of temporary workers across the U.S. between 1977 and 1997. We find evidence that all other things being equal, the THS share of employment is higher in states with more volatile industries. However, we also find that in a state with a relatively high degree of co-movement of industry output fluctuations, the use of temporary workers is lower, suggesting a reduced ability of THS agencies to enhance labor market flexibility in these states. Our finding suggests that THS agencies can operate more efficiently as an intermediary between client firms and workers in an environment in which industry output fluctuations do not coincide.

NOTES

[1]Under the Employee Retirement Income and Security Act, for a firm to receive a tax deduction on its contributions to its employee pension plan, the plan must cover at least 70 percent of non-highly compensated employees who worked 1,000 hours or more over the previous 12 months. Thus, many temporary workers may be excluded even if they work on a full-time basis (Houseman, 2001b).

[2]To accommodate to the data available from the BEA, based on the two-digit Standard Industrial Classification (SIC) system, we categorize SIC industries into 54 categories: 19 manufacturing industries, 13 service industries, eight transport and public utility industries, six finance and insurance industries, four mining industries, and one each for the construction, wholesale, retail trade, and agriculture industries. We excluded the agricultural industry in calculating the index. Statistics cited in Cohany (1998) indicate that THS agencies do not typically serve that industry.

[3]Note that we calculate the output volatility of each industry at national level instead of state level. This is because the state-level volatility of an industry might be influenced by the amount of THS services available in the state; this may not be appropriate to examine the role of the THS industry in facilitating the flexibility of volatile industries. For example, in a state where THS services are not readily available, firms may have to operate with low levels of temporary workers in their labor force. Without the flexibility of adjusting their labor force, some firms may find it difficult to survive, leaving only stable firms in the state. As a result, the industry output will be less volatile in the state with a lower THS industry share, which will contribute to the positive correlation between the volatility level and the THS industry share across states. The volatility will be relatively greater in a state with a higher THS industry share, not because the THS industry meets the needs of the firms with volatile output, but because the firms could not survive in other states with lower THS shares. By measuring an industry's volatility at national level, we alleviate this problem to the extent that industry composition is determined exogenously.

[4]THS industry share is not included in the calculation of volatility index.

[5]In the CBP, before 1987 the SIC code for the THS sector is 7362; after 1987, it is SIC7363. The 1987 revision to the Standard Industrial Classification System (SIC) expanded the Temporary Help Supply Services industry (7362) to a slightly broader aggregate, Personnel Supply Services (7363). To the degree that this expansion is proportional across states, it is absorbed by year effects. We acknowledge the Center for Governmental Studies at Northern Illinois University for providing the CBP with supplemented data.

[6]While the CBP data do not distinguish between temporary and permanent employees of THS establishments, the overwhelming majority of THS employees are temporary workers. For example, Manpower Inc. has approximately 22,400 staff employees (1.2 percent of its total work force), who oversaw the placement of 1.9 million temporary workers in 2001. (These numbers are based on data available at www.manpower.com/mpcom/index.jsp.)

[7]Results presented here are from robust regressions as suggested by Li (1985). This method takes account of the effects of outliers by giving them a smaller weight.

[8]We also performed regressions including a variable that measures the rate of inter-state migration, since it is possible that newly arrived residents may be more likely to enter the temporary labor force. The migration measure is based on the share of respondents in the CPS datasets that indicated they lived in a different state a year prior to their interview. The variable is only available from 1982 to 1997, so we run the regressions for that limited period. We found that the share of recently migrated population was positively associated with the level of THS employment, while our key results regarding the volatility index remained qualitatively the same.

REFERENCES

Abraham, Katharine G., and Susan K. Taylor, 1996, "Firms' use of outside contractors: Theory and evidence," *Journal of Labor Economics*, Vol. 14, July, pp. 394–424.

Autor, David, 2003, "Outsourcing at will: Unjust dismissal doctrine and the growth of temporary help employment," *Journal of Labor Economics*, Vol. 21, No. 1, January, pp. 1–42.

_____, 2001, " Why do temporary help firms provide free general skills training?," *Quarterly Journal of Economics*, Vol. 116, No. 4, November, pp. 1409–1448.

Clinton, Angela, 1997, "Flexible labor: Restructuring the American work force," *Labor Monthly Review*, Vol. 120, No. 8, August, pp. 3–17.

Cohany, Sharon R., 1998, "Workers in alternative employment arrangements: A second look," *Monthly Labor Review*, Vol. 121, No. 11, November, pp. 3–21.

Conroy, Michael L., 1975, "The concept and measurement of regional industrial diversification," *Southern Economic Journal*, Vol. 41, pp. 495–505.

Diamond, Charles A., and Curtis J. Simon, 1990, "Industrial specialization and the returns to labor," *Journal of Labor Economics*, Vol. 8, No. 2, pp. 175–201.

Estẽvao, Marcello, and Saul Lach, 1999, "Measuring temporary labor outsourcing in U.S. manufacturing," Board of Governors of the Federal Reserve System, working paper, October.

Golden, Lonnie, 1996, "The expansion of temporary help employment in the US, 1982–1992: A test of alternative economic explanations," *Applied Economics*, Vol. 28, pp. 1127–1141.

Houseman, Susan N., 2001a, "Why employers use flexible staffing arrangements: Evidence from an establishment survey," *Industrial and Labor Relations Review*, Vol. 55, No. 1, October.

_____, 2001b, " The benefits implication of recent trends in flexible staffing arrangements," University of Pennsylvania, Wharton School, Pension Research Council, working paper, No. PRC WP 2001-19.

Li, G., 1985, "Robust regression," in *Exploring Data Tables, Trends, and Shapes*, D. C. Hoaglin, F. Mosteller, and J. W. Tukey (eds.), New York: John Wiley & Sons, pp. 281–340.

Otto, Maria Ward, 1999, "Temporary employment and the natural rate of unemployment," Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series, Washington DC.

Polivka, Anne E., 1996, "A profile of contingent workers," *Monthly Labor Review*, October.

Rogers, Jackie Krases, 2000, *Temps: The Many Faces of the Changing Workplace*, Ithaca, NY, and London: Cornell University Press.

Segal, Lewis M., and Daniel G. Sullivan, 1998, "Wage differentials for temporary service work: Evidence from administrative data," Federal Reserve Bank of Chicago, working paper, No. 98-23.

_____, 1997, "The growth of temporary services work," *Journal of Economic Perspectives*, Spring, pp. 117–136.

_____, 1995, "The temporary labor force," *Economic Perspectives*, Federal Reserve Bank of Chicago, March–April, pp. 2–19.

Staffing Industry Analysis Inc., 2002, *Staffing Industry Report*, Vol. 13, No. 14, July.

U.S. Department of Commerce, Bureau of the Census, 1977–97, *County Business Patterns*, Washington, DC.

_____, 1992, *Enterprise Statistics*, Washington, DC.

The optimal price of money

Pedro Teles

Introduction and summary

One of the basic monetary policy issues facing the monopolist supplier of currency is what price to charge for its use. The price paid for the use of currency, by households or firms, is the foregone interest on less liquid, but riskless, assets such as short-term government bonds. Thus, the question of what price to charge for the use of currency is identified with the question of what is the optimal nominal interest rate.

According to Friedman (1969), monetary policy ought to be conducted so that the resulting nominal interest on short-term, less liquid assets is zero. The argument for the Friedman rule is very simple: Since the cost of supplying money is negligible,[1] the price charged for its use should also be very close to zero.

The first best argument of Friedman (1969) was challenged by Phelps (1973) on the basis that a positive nominal interest rate generates tax revenues for the government. According to Phelps (1973), since the alternative sources of revenue also create distortions, liquidity should be taxed like any other good. This public finance argument motivated a literature on the optimal inflation tax in a second-best environment, where the government is constrained to finance exogenous government expenditures by recourse to distortionary taxes. Somewhat surprisingly, the recent literature on the optimal inflation tax has argued that, even in a second-best environment, it is optimal not to use the inflation tax, so that the Friedman rule is still optimal. Why is this the case? Why shouldn't liquidity be taxed like any other good, as argued by Phelps (1973)?

In this article, I review some of the results obtained in the literature on the optimality of the Friedman rule. I base the analysis on Correia and Teles (1996, 1999) and De Fiore and Teles (2003), which have built on work by Kimbrough (1986), Guidotti and Végh (1993), and Chari, Christiano, and Kehoe (1996), among others.

I start by analyzing a simple environment where liquidity services are modeled as a final good, so that agents gain utility from consumption, leisure, and real balances, measured by the stock of money deflated by the price level. This is the context in which the argument of Phelps (1973) was made. According to Phelps, an application of the Ramsey (1927) principles of taxation of final goods, would mean that tax distortions should be distributed across goods, including liquidity services. Since the public finance principles, such as Ramsey (1927), were applied to costly goods, I allow for the possibility that money is costly to supply. I assume that the utility function satisfies the conditions for uniform taxation of final goods, established by Atkinson and Stiglitz (1972). In that case it is optimal to tax money, at the same proportionate rate as the consumption good. Thus, the price charged for the use of money, the nominal interest rate, *int*, should be equal to the cost of producing real balances, c, marked up by the optimal common tax rate, τ^*, on real balances and consumption,

$$int^* = c(1 + \tau^*).$$

As the cost of producing money, c, is reduced, so is the optimal price charged for the use of money, int^*. When the cost is zero, $c = 0$, the optimal nominal interest rate is also zero,

$$int^* = 0.$$

Thus, even if the optimal proportionate tax on money is positive and relatively high, because the production costs of money are very small, the optimal price charged for money and therefore the implicit unit tax

Pedro Teles is a senior economist at the Federal Reserve Bank of Chicago.

may also be very small. In this case, it is clear that the reason for the optimality of the Friedman rule is the fact that money is costless.

Since real balances measure the purchasing power of money, it is more appropriate to use, as its measure, the stock of money deflated by the price level gross of consumption taxes, rather than net of these taxes. The reason for this is that the consumption taxes are typically paid using the same means of payment as that used to purchase the consumption goods. A small modification of the model described above considers this measure of real balances. The fact that money balances are deflated by the price level gross of taxes implies that the price paid for the use of real balances is now the nominal interest rate marked up by the consumption tax. Under the conditions for uniform taxation, in order to guarantee that real money is taxed at the same rate as consumption goods, the nominal interest rate ought to be equal to the production cost of real balances c,

$$int^* = c.$$

If the production cost c is negligible, then the nominal interest rate should be zero. Thus, also in this case, money is optimally taxed at a positive proportionate rate. However, the total price charged for money when the cost of producing money is zero is still zero,

$$int^*(1 + \tau^*) = 0.$$

Again in this case, the Friedman rule is optimal because of the assumption of a negligible production cost of money.

In the examples just described, liquidity was treated as a final good like any other consumption good. In reality, liquidity is valued because it reduces transaction costs. Modeling money as an input in the production of transactions, rather than as an argument in the utility function, has implications for the optimal inflation tax when money is costly to produce. Under the assumption that money is costly, if the transactions technology is constant returns to scale, real balances should not be taxed.[2] Thus, the optimal tax rate on real balances is

$$\tau^* = 0.$$

This is in the spirit of Diamond and Mirrlees' (1971) taxation rules, whereby it is not optimal to tax intermediate goods when the technology is homogenous of degree one. If instead, the degree of homogeneity of the transactions technology is different from one, as in the case of the transactions technology proposed by Baumol (1952) and Tobin (1956), then it is optimal to set a non-zero tax on the use of real balances,

$$\tau^* \lesseqgtr 0.$$

The optimal proportionate tax or subsidy does not approach zero as the cost of producing money becomes arbitrarily small. However, in the limit, when $c = 0$, the price of using money, that is, the nominal interest rate marked up by the consumption tax, $int^*(1 + \tau^c)$, is zero,

$$int^*(1 + \tau^c) = c\,(1 + \tau^*) = 0.$$

The Friedman rule is optimal. Thus, in this environment as well, it is the costless nature of money that justifies not taxing real balances. I review these results based on Correia and Teles (1996, 1999).

The analysis in this article compares, in welfare terms, consumption taxes to the inflation tax and leaves out income taxes. The reason for this is that, under reasonable assumptions on the transactions technology, consumption and income taxes are equivalent tax instruments, and so the result on the optimal inflation tax is unchanged whether one or the other alternative tax is considered. That is not the case when one uses the standard specification of the transactions technology, first proposed by Kimbrough (1986). Consequently, the issue of which alternative tax instrument one considers has received some attention in the literature. When the alternative tax is an income tax, the Friedman rule is optimal, while when the alternative is a consumption tax, the conditions for the optimality of the Friedman rule are more restrictive. Mulligan and Sala-i-Martin (1997) used this fact to argue for the fragility of the Friedman rule. I review their claim, which is assessed in De Fiore and Teles (2003).

The policy implications from the analysis in this article should be taken with some caution, since the analysis abstracts from the role of monetary policy as stabilization policy, justified by the presence of nominal rigidities that are assumed away in the analysis. In models with those frictions, although there are simple structures where the Friedman rule is still optimal (see Correia, Nicolini, and Teles, 2001), in more complex staggered price-setting environments, the optimality of the Friedman rule is lost. Nevertheless, the optimal inflation rate is still a very low number. Another aspect of monetary policy that this analysis abstracts from is the issue of commitment. The assumption here is that the policymaker can commit to future policy. If that is not the case, the policy suggestions in this article will not be of much use.

A simple model of liquidity as a final good

The first model I consider is a simple money-in-the-utility-function model. In such models, agents use real balances because they provide utility directly. This assumption is useful in the context of the analysis in this article to assess the public finance argument, originally made by Phelps (1973), that liquidity should be taxed like any other good.

The preferences of the households depend on consumption, leisure (defined here as time not devoted to the production of the consumption good), and real balances. In a first version of the model, I define real balances as the nominal balances deflated by the price level net of consumption taxes. The goods are produced with time and, for the sake of understanding the implications of money being a costly good, there is also a time cost of real balances.[3] The government must finance exogenous expenditures with either consumption taxes or the inflation tax. A positive inflation tax is levied whenever the price charged for the use of money is higher than the cost of producing it, that is, when the nominal interest rate is higher than the time cost of producing real balances. When that cost is zero and the interest rate is also zero, the Friedman rule is followed. When the cost is positive, a modified Friedman rule, which takes into account that money is costly, sets the nominal interest rate equal to the cost of real balances.

In this model the nominal interest rate creates a distortion between real balances and leisure when it differs from the cost of producing real balances. A non-zero consumption tax creates a distortion between consumption and leisure. In this model where real balances are a final good, a direct application of the Ramsey (1927) principles of taxation would suggest that real balances ought to be taxed like any other good. Indeed, under the conditions on preferences established by Atkinson and Stiglitz (1972), the two goods, consumption and money, should be taxed at the same proportionate rate. Therefore, under those conditions, the nominal interest rate should be equal to the production cost of money marked up by the proportionate tax levied on the consumption goods. This means that even for a very small cost of producing money, the modified Friedman rule is not exactly optimal. It is approximately optimal, though.

The Friedman rule is optimal in the limit case where the cost of supplying money is exactly zero. As the cost of producing money approaches zero, the consumption tax converges to a finite and strictly positive number, and thus the optimal price charged for the use of money converges to the production cost, that is, zero. In this case, it is clear that the optimality of the Friedman rule hinges on the assumption that money is costless. The formal analysis of this problem is described in box 1.

Money is deflated by the price level gross of consumption taxes

Above, I assumed that liquidity services were represented, as a final good, by the stock of nominal money deflated by the price level net of taxes. However, if consumption taxes are paid with money, the liquidity services of money are more appropriately described by the stock of money deflated by the price level gross of consumption taxes. What are the implications of considering this measure of real balances?

If liquidity services are measured by money deflated by the price level gross of consumption taxes, money is implicitly taxed at the same rate as consumption, and so the cost of using money is no longer the nominal interest rate, but rather the interest rate marked up by the consumption tax. The relative price of real balances in units of time is $i_t(1 + \tau_{ct})$, while the relative price of consumption in units of time is $(1 + \tau_{ct})$. Under the conditions for uniform taxation of Atkinson and Stiglitz (1972), the optimal nominal interest rate is equal to the cost of supplying real balances,[4]

$$i_t = \alpha.$$

Does this mean that the Friedman rule is optimal? Not really. In this context, a modified Friedman rule should take into account the implicit taxation of money, resulting from the need to use money to pay taxes. The modified Friedman rule is such that the total cost of using money equals the cost of supplying it,

$$i_t(1 + \tau_{ct}) = \alpha.$$

Thus, in order for the modified Friedman rule to hold, the nominal interest rate would have to include a subsidy to money at the same rate as the consumption tax that would compensate for the implicit taxation of real balances.

Under the conditions for uniform taxation, this policy is not optimal. However, as the cost of supplying money approaches zero, the two policies coincide. The optimal policy is the Friedman rule of a zero nominal interest rate. Again, in this case the Friedman rule is optimal because money has a zero cost of production. The Ramsey problem in this environment is formalized in box 2.

A monetary model with a transactions technology

The money-in-the-utility-function models analyzed in the previous sections can be interpreted as equivalent representations of models where money reduces the transactions costs that households have

BOX 1

The Ramsey problem in a money-in-the-utility-function model

In this model with money in the utility function, preferences depend on consumption c_t, real balances $\tilde{m}_t = \frac{M_t}{P_t}$, where P_t is the price level net of consumption taxes, and time not devoted to producing the consumption good that I call leisure, h_t^v,

$$1)\quad \sum_{t=0}^{\infty} \beta^t V(c_t, \frac{M_t}{P_t}, h_t^v).$$

The technology to produce consumption uses time only and is linear with a unitary coefficient.

The representative household chooses a sequence $\{c_t, h_t^v, M_t, B_t\}_{t=0}^{\infty}$, where B_t are nominal securities that pay $(1 + i_t)B_t$ units of money in period $t + 1$, that satisfies the budget constraint and maximizes utility in equation 1, given a sequence of prices, $\{P_t, i_t\}_{t=0}^{\infty}$ and initial nominal wealth $W_0 \equiv M_{-1} + (1 + i_{-1})B_{-1}$. For simplicity, I assume that the initial wealth is zero, $W_0 = 0$. The budget constraint is described by the following sequence:

$$2)\quad M_{t+1} + B_{t+1} \le M_t - (1+\tau_{ct})P_t c_t + P_t(1-h_t^v) + (1+i_t)B_t,\ t \ge 0$$
$$M_0 + B_0 \le W_0$$

together with a no-Ponzi games condition. The variable τ_{ct} is the consumption tax rate.

The government finances an exogenous sequence of government expenditures, $\{g_t\}$, by setting tax rates on the consumption good, $\{\tau_{ct}\}$, as well as the nominal interest rates, $\{i_t\}$. The resource constraints in this economy are given by

$$3)\quad c_t + g_t \le 1 - h_t^v - \alpha\tilde{m}_t,\ t \ge 0,$$

where α is the cost in units of time of supplying one unit of real money. It is a standard assumption in the literature that this cost is zero, $\alpha = 0$.

The intertemporal budget constraint for consumers can be written as

$$4)\quad \sum_{t=0}^{\infty} \frac{Q_t P_t}{Q_0 P_0}(1+\tau_{ct})c_t + \sum_{t=0}^{\infty} \frac{Q_t P_t}{Q_0 P_0} i_t \tilde{m}_t \le \sum_{t=0}^{\infty} \frac{Q_t P_t}{Q_0 P_0}(1-h_t^v),$$

where $Q_t = \frac{1}{(1+i_0)\ldots(1+i_t)}$, $t \ge 0$. Maximizing equation 1 subject to equation 4, I obtain the following marginal conditions:

$$5)\quad \frac{V_c(t)}{V_{h^v}(t)} = 1 + \tau_{ct},\ t \ge 0$$

$$6)\quad \frac{V_{\tilde{m}}(t)}{V_c(t)} = \frac{i_t}{(1+\tau_{ct})},\ t \ge 0$$

$$7)\quad \frac{\beta^t V_{h^v}(t)}{V_{h^v}(0)} = \frac{Q_t P_t}{Q_0 P_0},\ t \ge 0.$$

The marginal conditions 5–7, the budget constraint, 4 satisfied with equality, and the resource constraints, 3, determine the set of feasible and implementable allocations, $\{c_t, h_t^v, \tilde{m}_t\}_{t=0}^{\infty}$, intertemporal prices $\left\{\frac{Q_t P_t}{Q_0 P_0}\right\}_{t=0}^{\infty}$, and taxes $\{\tau_{ct}, i_t\}_{t=0}^{\infty}$. This is the set of competitive equilibriums, such that the government finances exogenous government expenditures with consumption and inflation taxes. The government solves a Ramsey problem, by choosing in this set the path for the quantities, prices, and taxes that maximizes welfare, thus minimizing the excess burden of taxation.

The two intratemporal marginal conditions 5 and 6 and the resource constraint 3 determine the quantities of consumption, leisure, and real balances in each period $t \ge 0$ as functions of the taxes.[1] Once $\{c_t, h_t^v, \tilde{m}_t\}_{t=0}^{\infty}$ are determined as functions of the taxes $\{\tau_{ct}, i_t\}_{t=0}^{\infty}$, I can use condition 7 to determine the path of the intertemporal prices, $\left\{\frac{Q_t P_t}{Q_0 Q_0}\right\}_{t=0}^{\infty}$ as functions of the taxes $\{\tau_{ct}, i_t\}_{t=0}^{\infty}$. The paths of taxes must satisfy the government's budget constraint, which can be obtained from the households' budget constraint, 4 with equality, and the resource constraints. This strategy of solving the system of competitive equilibrium equations is the dual approach. Because the system is linear in the taxes and prices, a primal approach is more efficient, where the taxes and prices are expressed as functions of the quantities, and substituted in the households' budget constraint.

Thus, I substitute the tax rates and prices, $\left\{\tau_{ct}, i_t, \frac{Q_t P_t}{Q_0 P_0}\right\}_{t=0}^{\infty}$ using conditions 5–7 into the budget constraint, 4 satisfied with equality, and obtain the implementability condition

$$8)\quad \sum_{t=0}^{\infty}\beta^t[V_c(t)c_t + V_{\tilde{m}}(t)\tilde{m}_t - V_{h^v}(t)(1-h_t^v)]=0.$$

The Ramsey problem will then be simplified to consist of the choice of the path of quantities, $\{c_t, h_t^v, \tilde{m}_t\}_{t=0}^{\infty}$, that satisfies the implementability condition 8 and the resource constraint 3 and maximizes welfare. The taxes and prices that decentralize the optimal solution can then be obtained from equations 5–7. The following are first order conditions:

$$9)\quad V_c(t) + \psi[V_c + V_{cc}(t)c_t - V_{h^v c}(t)(1-h_t) + V_{\tilde{m}c}(t)\tilde{m}_t] = \lambda_t,\ t \geq 0$$

$$10)\quad V_{h^v}(t) + \psi[V_{h^v}(t) + V_{ch^v}(t)c_t - V_{h^v h^v}(t)(1-h_t) + V_{\tilde{m}h^v}(t)\tilde{m}_t] = \lambda_t,\ t \geq 0$$

$$11)\quad V_{\tilde{m}}(t) + \psi[V_{\tilde{m}}(t) + V_{c\tilde{m}}(t)c_t - V_{h^v\tilde{m}}t)(1-h_t) + V_{\tilde{m}\tilde{m}}(t)\tilde{m}_t] = \alpha\lambda_t,\ t \geq 0,$$

where ψ and $\beta^t\lambda_t$ are the multipliers of the implementability constraint and the time t resource constraint, respectively.

Suppose the utility function is additively separable in leisure and homogeneous in consumption and real balances,[2] so that it can be written as

$$12)\quad V(c, \tilde{m}, h) = u(c, \tilde{m}) + v(h^v),$$

where u is homogeneous of degree k. Then the first order conditions of the Ramsey problem in equations 9 and 10 become

$$13)\quad V_c(t)[1+\psi(1+k)] = \lambda_t$$

$$14)\quad V_{\tilde{m}}(t)[1+\psi(1+k)] = \alpha\lambda_t.$$

From these, I obtain $\frac{V_{\tilde{m}}(t)}{V_c(t)} = \alpha$. Thus, the optimal policy will not distort the marginal choice between consumption and real balances.[3] The way to decentralize this solution is to set the same proportionate tax on consumption and money. Since, from equation 6, the relative price of $\tilde{m}$ in units of consumption is $\frac{i_t}{(1+\tau_{ct})}$, the optimal interest rate is $i_t = (1+\tau_{ct})\alpha$, so that it imposes a tax on real balances, at the same rate as the consumption tax.

In this context, where there is a cost of producing real balances, it makes sense to consider a modified Friedman rule that takes into account the production cost of money and corresponds to a zero tax on money. According to that modified rule, the nominal interest rate should equal the cost of producing real balances, $i_t = \alpha$. The modified Friedman rule is not optimal. This is true for any $\alpha > 0$ since the tax rate on consumption is bounded away from zero.

Consider a sequence of problems where the cost of supplying real balances, α, approaches zero. Since the optimal tax rate on consumption is bounded above, as the cost of supplying real balances becomes arbitrarily low, the optimal interest rate approaches zero. Thus, in the limit, the Friedman rule is optimal. In this case, it is clear that the reason the Friedman rule is optimal is the standard assumption of a zero cost of producing real balances.

---

[1]If the taxes and the government expenditures are constant over time, $\tau_{ct} = \tau_c$, $i_t = i$, and $g_t = g$, then the allocation will be stationary. This steady state will be characterized by $\frac{1}{\beta} = (1+i)\frac{P_t}{P_{t+1}}, t \geq 0$, so that inflation will be constant as well. In this stationary economy, inflation will be equal to the growth rate of money supply. In order for the nominal interest rate to be equal to zero, so that the Friedman rule is followed, it must be the case that the growth rate of money supply is negative and equal to $\beta - 1$.

[2]The conditions for uniform taxation of Atkinson and Stiglitz (1972) are separability of leisure and homotheticity in the consumption goods.

[3]In one specification of $u(c, \tilde{m}) = \frac{c^{1-\sigma}}{1-\sigma} + B\frac{\tilde{m}^{1-\sigma}}{1-\sigma}$, homogeneity corresponds to equal elasticity. In this case, where the goods have the same price elasticity, the tax rates ought to be the same.

BOX 2

The Ramsey problem with real balances measured by money deflated by the price level gross of taxes

The utility function is

$$15)\quad \sum_{t=0}^{\infty}\beta^t V\left(c_t, \frac{M_t}{(1+\tau_{ct})P_t}, h_t^v\right)$$

and the resource constraints are given by

$$16)\quad c_t + g \le 1 - h_t^v - \alpha m_t,\ t \ge 0,$$

where α is the cost in units of time of supplying one unit of real money, $m_t \equiv \frac{M_t}{(1+\tau_{ct})P_t}$.

The representative household maximizes utility in equation 15, subject to the intertemporal budget constraint

$$17)\quad \sum_{t=0}^{\infty}\frac{Q_t P_t}{P_0}(1+\tau_{ct})c_t + \sum_{t=0}^{\infty}\frac{Q_t P_t}{P_0} i_t (1+\tau_{ct}) m_t \le \sum_{t=0}^{\infty}\frac{Q_t P_t}{P_0}(1-h_t^v).$$

The marginal conditions are

$$18)\quad \frac{V_c(t)}{V_{h^v}(t)} = 1+\tau_{ct},\ t \ge 0$$

$$19)\quad \frac{V_m(t)}{V_c(t)} = i_t,\ t \ge 0$$

$$20)\quad \frac{\beta^t V_{h^v}(t)}{V_{h^v}(0)} = \frac{Q_t P_t}{Q_0 P_0},\ t \ge 0.$$

Proceeding as before, I obtain the implementability condition

$$21)\quad \sum_{t=0}^{\infty}\beta^t [V_c(t)c_t + V_m(t)m_t - V_{h^v}(t)(1-h_t^v)] = 0.$$

The two Ramsey problems are identical once $\tilde{m}_t$ is replaced for m_t. Thus, when the utility function is additively separable in leisure and homogeneous in consumption and real balances, we have

$$\frac{V_m(t)}{V_c(t)} = \alpha.$$

As before, the optimal fiscal policy will not distort the marginal choice between consumption and real balances. However, in this case, since the relative price of real balances in terms of consumption is i_t, the optimal solution will be decentralized with

$$i_t = \alpha,$$

so that the nominal interest rate does not include a tax or a subsidy on money. Money is being taxed implicitly at the same rate as consumption.

As the cost of supplying real balances becomes arbitrarily low, the optimal interest rate approaches zero. In the limit, the price charged for the use of money is zero. The Friedman rule is optimal in that limiting case.

to incur. That interpretation is the common justification for the assumption that real balances are a final good. In this section, I analyze a standard model with a transactions technology, derive the equivalent money-in-the-utility-function model, and show that the restrictions imposed by the transactions technology structure would have implications for the optimal inflation tax if money was a costly good. In the models above, the optimal policy imposed the same proportionate distortion between real balances and leisure as between consumption and leisure; however, when money is modeled as an input into a constant returns to scale transactions technology, it is no longer optimal to distort the marginal choice between real balances and leisure, so that a modified Friedman rule is optimal. The latter result is an application of the optimal taxation rules of Diamond and Mirrlees (1971), whereby intermediate goods should not be taxed when consumption taxes are available and the technology is constant returns to scale.

Since the monetary aggregate that facilitates transactions is, by assumption, the stock of money deflated by the price level gross of consumption taxes, real balances are implicitly taxed at the consumption tax rate, so that the relative price of money in terms of leisure is $i_t(1 + \tau_{ct})$. The relative price of consumption is $(1 + \tau_{ct})$. When the transactions technology is constant returns to scale, the optimal policy is to set the price of money equal to its cost,

$$i_t(1 + \tau_{ct}) = \alpha,$$

so that the optimal nominal interest rate includes a subsidy at the consumption tax rate. A modified Friedman rule, which takes into account both the cost of producing real balances and the implicit tax on real balances resulting from the need to use money to pay the consumption taxes, is optimal.

The principle that when there are consumption taxes it is not optimal to tax intermediate goods only holds if the technology is constant returns to scale, and, therefore, there are no implicit profits. Under the assumption that money is costly, if the transactions technology is not constant returns to scale, then the modified Friedman rule will no longer be optimal. In this case there are positive or negative implicit profits, introducing a trade-off between the lump-sum taxation of profits and the production distortions. In order to reduce profits, it will be optimal to either tax or subsidize money, depending on the degree of homogeneity.

The optimal policy is

$$22)\quad i_t = \frac{\alpha}{(1+\tau_{ct})}\left[\frac{1}{1+\frac{\psi U_h(t)[k-1]}{\lambda_t}}\right],$$

where ψ is the multiplier of the implementability condition, equation 33 in box 3, measuring the excess burden of taxes; λ_t is the multiplier of the resource constraint at time t; $U_h(t)$ is the marginal utility of leisure; and k is the degree of homogeneity of the transactions technology.[5] Clearly, if the transactions technology is constant returns to scale, so that $k = 1$, the modified Friedman rule is optimal. If $k > 1$, money should be subsidized, and if $k < 1$, money should be taxed. As α approaches zero, the Friedman rule is optimal for any value of the degree of homogeneity of the transactions technology. Thus, the modified Friedman rule is not generally optimal; it is optimal only for a zero cost of producing money.

Even if the two models, money in the utility function or transactions technology, give disparate results on the optimal inflation tax when money is costly, the limiting result, when the cost of producing money is zero, is the same. The Friedman rule is optimal, independent of the modeling assumption, because money is costless to produce.

The result that in models with transactions technologies it is optimal not to tax costless money was first obtained by Kimbrough (1986) and then extended by Chari, Christiano, and Kehoe (1996) and Correia and Teles (1996). The public finance exercise in these last two papers was a comparison between the income and inflation taxes. If instead the option is between the inflation tax and a consumption tax, other issues arise concerning the specification of the transactions technology. I discuss these issues, addressed by De Fiore and Teles (2003), in the next section.

The Ramsey problem in this section is formalized and solved in box 3.

How do consumption taxes affect the transactions technology?

In the previous section, I showed that, when the cost of producing money is zero, the Friedman rule is optimal for transactions technologies of any degree. This is the same result that Correia and Teles (1996) obtained in comparing the inflation tax with an income tax. In the set-up described above,[6] the consumption and income taxes are equivalent fiscal instruments. This means that the allocations that can be implemented are the same with any of the two taxes, so that the optimal inflation tax does not depend on which tax is considered. This result is in contrast with recent literature, in particular Mulligan and Sala-i-Martin (1997), who argue that the optimality of the Friedman rule is a fragile result because it depends on the alternative tax instrument. In this section, I clarify this point, based on De Fiore and Teles (2003).

Under the standard specification of the transactions technology, as originally proposed by Kimbrough (1986) and later used by Guidotti and Végh (1993) and Mulligan and Sala-i-Martin (1997), among others, the consumption and income taxes are, indeed, not equivalent fiscal instruments. Moreover, when the alternative tax instrument is the consumption tax, the conditions for the Friedman rule to be optimal are more restrictive. The reason for these contrasting results is that the standard specification of the transactions technology does not impose that money be unit elastic with respect to the price level gross of consumption taxes, as I assumed in equation 24 in box 3,

$$s_t = l\left(c_t, \frac{M_t}{(1+\tau_{ct})P_t}\right).$$

The transactions technology specified by Kimbrough (1986) is the following:

$$s_t = l\left(c_t(1+\tau_{ct}), \frac{M_t}{P_t}\right).$$

If the function l is homogeneous, then it can be written as

BOX 3

The Ramsey problem in a transactions technology model

In a monetary model with a transactions technology, the preferences of the representative household depend only on consumption and leisure, where leisure does not include the time used for transactions. They are given by

$$23)\quad \sum_{t=0}^{\infty}\beta^t U(c_t,h_t),$$

where U is an increasing concave function, c_t are consumption goods, and h_t is leisure at time t. The households supply labor $1-h_t-s_t$, where s_t is time spent in transactions.

Transactions are costly since they require time that could otherwise be used for production. The amount of time devoted to transactions increases with consumption, c_t, and decreases with real money balances, $m_t=\frac{M}{(1+\tau_{ct})P_t}$, where P_t is the price of the good before taxes, according to the following transactions technology:

$$24)\quad s_t \geq l\left(c_t,\frac{M_t}{(1+\tau_{ct})P_t}\right).$$

According to this transactions technology, money is unit elastic with respect to the price level gross of consumption taxes. In addition to standard assumptions to ensure that the problem is concave, it is assumed that the function l is homogeneous of degree $k \geq 0$.

The budget constraints of the households can be written as:

$$25)\quad \sum_{t=0}^{\infty}\frac{Q_tP_t}{Q_0P_0}(1+\tau_{ct})c_t+\sum_{t=0}^{\infty}\frac{Q_tP_t}{Q_0P_0}(1+\tau_{ct})i_tm_t \leq \sum_{t=0}^{\infty}\frac{Q_tP_t}{Q_0P_0}(1-h_t-l(c_t,m_t)),$$

where $Q_t=\frac{1}{(1+i_0)\ldots(1+i_t)}$.

The resource constraints are

$$26)\quad c_t+g_t \leq 1-h_t-l(c_t,m_t)-\alpha m_t.$$

This model can be written as an equivalent money-in-the-utility-function model by defining h_t^v as the total time used for leisure and transactions, $h_t^v \equiv h_t+s_t$.

The model can thus be written in the form presented in the last section. The preferences, the resource constraints, and the budget constraint are given by the following expressions:

$$27)\quad \sum_{t=0}^{\infty}\beta^t V(c_t,h_t^v,m_t)=\sum_{t=0}^{\infty}\beta^t U(c_t,h_t^v-l(c_t,m_t))$$

$$28)\quad c_t+g \leq 1-h_t^v-\alpha m_t$$

$$29)\quad \sum_{t=0}^{\infty}\frac{Q_tP_t}{Q_0P_0}(1+\tau_{ct})c_t+\sum_{t=0}^{\infty}\frac{Q_tP_t}{Q_0P_0}(1+\tau_{ct})i_tm_t \leq \sum_{t=0}^{\infty}\frac{Q_tP_t}{Q_0P_0}(1-h_t^v).$$

From equation 27, it becomes clear that the assumptions above that the utility function is separable in h_t^v and homogeneous in consumption and real balances are not easily justifiable.

The private problem is defined by the maximization of condition 27, subject to condition 29. The households' problem must satisfy the following marginal conditions:

$$30)\quad \frac{V_c(t)}{V_{h^v}(t)}=\frac{U_c(t)-U_h(t)l_c(t)}{U_h(t)}=1+\tau_{ct},\ t\geq 0$$

$$31)\quad \frac{V_m(t)}{V_c(t)}=-\frac{U_h(t)}{U_c(t)-U_h(t)l_c(t)}l_m(t)=i_t,\ t\geq 0$$

$$32)\quad \frac{\beta^t V_{h^v}(t)}{V_{h^v}(0)}=\frac{Q_tP_t}{Q_0P_0},\ t\geq 0.$$

Conditions 30–32, 29 with equality, and 28, determine the set of feasible and implementable allocations, $\{c_t,h_t,m_t\}_{t=0}^{\infty}$, prices $\left\{\frac{Q_tP_t}{Q_0P_0}\right\}_{t=0}^{\infty}$, and taxes $\{\tau_{ct},i_t\}_{t=0}^{\infty}$.

Using the fact that $l(t)$ is homogeneous of degree k, so that $ks_t = l_c(c_t, m_t)\,c_t + l_m(c_t, m_t)m_t$, the implementability condition 21 in box 2 can be written as

$$33)\quad \sum_{t=0}^{\infty}\beta^t[U_c(t)-U_h(t)(1-h_t^v+kl(t))]=0.$$

BOX 3 (CONTINUED)

The Ramsey problem is, therefore, the choice of quantities, $\{c_t, h_t^v, m_t\}_{t=0}^{\infty}$, that maximize welfare, represented by the utility function (condition 27), and satisfy the implementability condition 33 and the resource constraints 28.

The first order conditions of this problem imply the following condition,

$$34)\quad \frac{V_m(t)}{V_{h^v}(t)} = -l_m(t) = \alpha \frac{1}{1+\frac{\psi U_h(t)[k-1]}{\lambda_t}},$$

where ψ and $\beta^t\lambda_t$ are the multipliers of the implementability condition and the resource constraints, respectively. Then, when $k = 1$, there is no distortion imposed between real money and leisure. If $k < 1$, money should be taxed, and if $k > 1$, money should be subsidized.

Since real money is implicitly taxed, because money is needed to pay taxes, the implementation of this solution requires that money be subsidized. Since the private problem has

$$35)\quad \frac{V_m(t)}{V_{h^v}(t)} = -l_m(t) = i_t(1+\tau_{ct}),\ t \geq 0,$$

when $k = 1$, the optimal policy is

$$36)\quad i_t = \frac{\alpha}{(1+\tau_{ct})}.$$

This is the modified Friedman rule, in this context where there is a cost of producing real balances and they are implicitly taxed.

Whether the modified Friedman rule is optimal or not, as α is made arbitrarily low the Friedman rule is optimal. If $k \neq 1$, there is a tax or a subsidy that is, in absolute value, bounded above and away from zero, as α becomes arbitrarily low.[1,2] So in this case too, it is the negligible cost of money that justifies the zero tax on money.

[1]As shown in Correia and Teles (1996), the term $\frac{\psi U_h(t)(k-1)}{\lambda_t}$ measures the marginal effect of real balances on the implicit profits in the production of transactions. By taxing or subsidizing real balances, the planner aims to reduce profits.

[2]If the transactions technology is Baumol–Tobin, so that $k = 0$, then it is optimal to set a positive tax on money balances.

$$37)\quad s_t = (1+\tau_{ct})^k l\left(c_t, \frac{M_t}{(1+\tau_{ct})P_t}\right),$$

where k is the degree of homogeneity. Notice that, under this specification and not under equation 24, for $k > 0$, it is possible to reduce time used for transactions without adjusting the real quantity of transactions measured in units of the consumption good, c_t, and without changing the real quantity of money required to buy those goods, $\frac{M_t}{(1+\tau_{ct})P_t}$. An extreme example of this is when c_t and $\frac{M_t}{(1+\tau_{ct})P_t}$ are kept constant, while setting $\tau_{ct} = -1$. As a result, transactions will be zero, $c_t(1 + \tau_{ct}) = 0$, and so time used for transactions will also be zero.

Under the standard specification of the transactions technology in equation 37, it may be optimal to use the consumption tax to reduce the volume of transactions and save on resources. In particular, when both consumption and income taxes are allowed, it is optimal, under certain conditions, to fully tax income and subsidize consumption in order to eliminate transaction costs. When this is so, the government performs the full volume of transactions on behalf of the private agents. When the taxes on income are excluded, then it may be optimal to set a positive inflation tax, so that the consumption tax may be lower, and it may be possible to save on the volume and cost of transactions.

Box 4 describes the formal solution of the Ramsey problem in this alternative environment.

Conclusion

From a Ramsey perspective on the optimum quantity of money, the optimality of the Friedman rule owes its robustness to the costless nature of money. In general, if money was a costly good, a positive price should be charged for its use, and this price in general should be distorted. It is not clear whether this distortion should involve subsidizing or taxing money, but still there should in general be one. As the cost of producing money becomes arbitrarily low, the proportionate distortion is in absolute value bounded above and away from zero. Thus, it is the costless nature of money that justifies the optimality of the Friedman rule.

BOX 4

The Ramsey problem with the standard specification of the transactions technology

The preferences in the equivalent money-in-the-utility-function model are

$$38)\quad \sum_{t=0}^{\infty}\beta^t V(c_t,(1+\tau_{ct}),m_t h_t^v)\equiv \sum_{t=0}^{\infty}\beta_t U\left(c_t,h_t^v-(1+\tau_{ct})^k l(c_t,m_t)\right).$$

The conditions of the private problem are the same as described by conditions 17 and 18–20 in box 2. The implementability condition and the resource constraint are also the same, respectively, as conditions 21 and 16. It is useful to write the marginal conditions of the private problem, conditions 18 and 19, as

$$39)\quad \frac{U_c(t)-U_h(t)(1+\tau_{ct})^k l_c(t)}{U_h(t)}=1+\tau_{ct},\ t\geq 0$$

and

$$40)\quad \frac{-U_h(t)(1+\tau_{ct})^k l_m(t)}{U_c(t)-U_h(t)(1+\tau_{ct})^k l_c(t)}=i_t,\ t\geq 0.$$

From these we have,

$$41)\quad -(1+\tau_{ct})^k l_m(t)=i_t(1+\tau_{ct}),\ t\geq 0.$$

The implementability condition can be written as:

$$42)\quad \sum_{t=0}^{\infty}\beta^t\left[U_c(t)-U_h(t)\left(1-h_t^v+(1+\tau_{ct})^k kl(t)\right)\right]=0.$$

The Ramsey problem is to maximize utility in condition 38, subject to conditions 16 and 42 and to the constraint that $\tau_{ct}=\tau(c_t,h_t^v,m_t)$ is defined implicitly by condition 39. The first order conditions imply

$$43)\quad -\{\lambda_t+\psi U_h(t)[k-1]\}\left[1+\tau_{ct})^k l_m(t)+k(1+\tau_{ct})^{k-1}\frac{\partial\tau_{ct}}{\partial m_t}l(t)\right]=\alpha\lambda_t,$$

since $\frac{\partial\tau_{ct}}{\partial h_t^v}=0$. The marginal effect of real balances on the consumption tax is

$$44)\quad \frac{\partial\tau_{ct}}{\partial m_t}=\frac{-(1+\tau_{ct})^k l_{cm}(t)}{\left[1+k(1+\tau_{ct})^{k-1}l_c(t)\right]}.$$

The second term on the left-hand side of equation 43, $k(1+\tau_{ct})^{k-1}\frac{\partial\tau_{ct}}{\partial m_t}l(t)$, is the impact on transactions time of a marginal increase in real balances through the effect on expenditures. If, at the Friedman rule, those effects are not zero, then the Friedman rule will no longer be optimal. That term is zero when either $k=0$, $l(t)=0$, or $l_{cm}(t)=0$. When it is not zero, even when $k=1$, it is not optimal to set the private benefit of money equal to the production cost, so that the result of zero taxation of intermediate goods of Diamond and Mirrlees (1971) does not apply. The reason is that, in this case, the technology is directly affected by the tax instruments.

NOTES

[1]The nominal production costs of currency as a percentage of its nominal value are approximately 0.12 percent. They are relatively high for small denomination bills (2.18 percent for $1 bills) but very low for higher denomination bills (less than 0.01 percent for $100 bills). The cost of coins is 0.94 percent.

[2]The transactions technology uses real balances and time to produce transactions measured by consumption.

[3]I assume a constant cost per unit of real balances. One rationale for this is the assumption that the production of real balances uses bills of different real denominations in fixed proportions and with constant returns, and that there is a constant time cost of producing bills of each real denomination. The second assumption would be more easily justified if the nominal denominations were indexed to the price level at zero cost. In reality, this indexation is costly.

[4]Notice that even if the measure of real balances is different, I maintain the assumption of a constant time cost per unit of real balances.

[5]The assumption that the transactions technology is homogenous of degree k means that when real balances and consumption are multiplied by λ, time used for transactions is multiplied by λ^k.

[6]The specification for the transactions technology in that section is as in De Fiore and Teles (2003).

REFERENCES

Atkinson, A. B., and Stiglitz, J. E., 1972, "The structure of indirect taxation and economic efficiency," *Journal of Public Economics*, Vol. 1, pp. 97–111.

Baumol, William J., 1952, "The transactions demand for cash: An inventory theoretic approach," *Quarterly Journal of Economics*, Vol. 66, pp. 545–556.

Chari, V. V., Lawrence J. Christiano, and Patrick J. Kehoe, 1996, "Optimality of the Friedman rule in economies with distorting taxes," *Journal of Monetary Economics*, Vol. 37, pp. 203–223.

Correia, Isabel, and Pedro Teles, 1999, "The optimal inflation tax," *Review of Economic Dynamics*, Vol. 2, pp. 325–346.

_____, 1996, "Is the Friedman rule optimal when money is an intermediate good?," *Journal of Monetary Economics*, Vol. 38, pp. 223–244.

Correia, Isabel, Juan Pablo Nicolini, and Pedro Teles, 2001, "Optimal fiscal and monetary policy: Equivalence results," Federal Reserve Bank of Chicago, mimeo.

De Fiore, Fiorella, and Pedro Teles, 2003, "The optimal mix of taxes on money, consumption, and income," *Journal of Monetary Economics*, Vol. 50, No. 4, also Federal Reserve Bank of Chicago, 2002, working paper, No. wp-2002-03.

Diamond, Peter A. and James A. Mirrlees, 1971, "Optimal taxation and public production," *American Economic Review*, Vol. 63, pp. 8–27, 261–268.

Friedman, Milton, 1969, "The optimum quantity of money," in *The Optimum Quantity of Money and Other Essays*, M. Friedman (ed.), Chicago: Aldine, pp. 1–50.

Guidotti, Pablo E., and Carlos A. Végh, 1993, "The optimal inflation tax when money reduces transactions costs," *Journal of Monetary Economics*, Vol. 31, pp. 189–205.

Kimbrough, Kent P., 1986, "The optimum quantity of money rule in the theory of public finance," *Journal of Monetary Economics*, Vol. 18, pp. 277–284.

Mulligan, Casey B., and Xavier X. Sala-i-Martin, 1997, "The optimum quantity of money: Theory and evidence," *Journal of Money, Credit, and Banking*, Vol. 29, No. 4, pp. 687–715.

Phelps, Edmund S., 1973, "Inflation in the theory of public finance," *Swedish Journal of Economics*, Vol. 75, pp. 37–54.

Ramsey, Frank P., 1927," A contribution to the theory of taxation," *Economic Journal*, Vol. 37, pp. 47–61.

Tobin, James, 1956, "The interest elasticity of the transactions demand for cash," *Review of Economics and Statistics*, Vol. 38, pp. 241–247.

Testing the Calvo model of sticky prices

Martin Eichenbaum and Jonas D. M. Fisher

Introduction and summary

A classic question in macroeconomics is: Why do changes in monetary policy affect aggregate economic activity? The answer to this question is of central importance to monetary policymakers. This is because policymakers require a convincing model of the monetary transmission mechanism in order to evaluate the consequences of alternative choices.

A key assumption in many models used to assess the effects of monetary policy is that the nominal prices of goods are "sticky." By this we mean that firms do not change their prices each period in response to the different shocks impacting on their environment. Models embodying this assumption typically have the property that policy actions that raise the money supply and lower short-term interest rates lead to expansions in aggregate economic activity. These types of models are increasingly being used by central banks around the world to help guide policymakers in setting monetary policy.

Different approaches to modeling sticky prices have been adopted in the literature. In one class of models, referred to as *time-dependent models*, the number of firms that change prices in any given period is specified exogenously. Classic models of this sort were developed by Taylor (1980) and Calvo (1983). Modern variants are now central elements of a large class of models.[1] A key feature of Calvo–Taylor pricing models is that forward-looking firms understand they will only periodically reoptimize prices. So, firms front load higher future expected real marginal costs into their current prices. They do this because they may not be able to raise prices when the higher marginal costs materialize. Similarly, to avoid declines in their relative prices, firms front load future inflation into the prices that they set. As typically formulated, these models often imply that deviations of economy-wide inflation from its long-run value depend primarily on current and expected changes in firms' real marginal costs.

In a different class of models, often referred to as *state-dependent pricing models*, the number of firms changing prices in any given period is determined endogenously. Dotsey, King, and Wolman (1999) model this phenomenon by assuming that firms pay a fixed cost when they change their price. In contrast, Burstein (2002) assumes that firms pay a fixed cost for changing price plans. Once they pay this cost, firms can choose not only their current price, but also a plan specifying an entire sequence of future prices. A key property of these models is that small and large changes in monetary policy have qualitatively different effects on aggregate economic activity.

While state-dependent models seem very promising (at least to us), they are substantially more difficult to work with than time-dependent models. In addition, the two classes of models generate similar results for many policy experiments that are relevant in moderate inflation economies such as the U.S.[2] For these reasons, modern variants of Taylor and Calvo models continue to play a central role in the analysis of monetary policy.

This article addresses the question: Are time-dependent models good models in an empirical sense? For concreteness, we focus on the Calvo sticky pricing model. In principle, there are a variety of ways to test this model. For example, one could embed it in a fully specified general equilibrium model of the economy. This would involve, among other things, modeling household labor and consumption decisions, credit markets, fiscal policy, and monetary policy. If in addition, one specified the nature of all the shocks impacting on the

Martin Eichenbaum is a professor of economics at Northwestern University, a research fellow at the National Bureau of Economic Research, and a consultant to the Federal Reserve Bank of Chicago. Jonas D. M. Fisher is a senior economist and economic advisor at the Federal Reserve Bank of Chicago.

economy, one could estimate and test the model using a variety of statistical methods like maximum likelihood.[3] Another strategy would be to assess the model's predictions for a particular shock, such as a disturbance to monetary policy or a shock to technology.[4]

Here, we focus on tests of the model using the econometric strategy pioneered by Hansen (1982) and Hansen and Singleton (1982) and applied to the Calvo model by Gali and Gertler (1999) and Eichenbaum and Fisher (2003). The idea is to exploit the fact that in any model incorporating Calvo pricing, certain restrictions must hold. One can test these restrictions, without making assumptions about other aspects of the economy. Of course, in the end, we need a fully specified model of the economy within which to assess the consequences of alternative policy. The approach that we discuss here has the advantage of focusing on the empirical plausibility of one key building block that could be an element of many models.

Our analysis proceeds as follows. In the next section, we summarize the Calvo model. Standard versions of the model assume that when firms reoptimize their price plans, the new plan takes effect immediately. As in Eichenbaum and Fisher (2003), we allow for the possibility that when firms reoptimize their price plans at time t, the new plan only goes into effect at time $t + \tau$, where $\tau \geq 0$ and a time period corresponds to a quarter. The standard Calvo model corresponds to the assumption that τ is equal to zero. By varying τ, we can vary the information set that firms have at their disposal when making new price decisions.

The following section discusses an econometric strategy for estimating and testing the model, taking into account the possibility of measurement error in the variables of interest, particularly inflation. Then, we discuss the four measures of inflation that we use in our empirical analysis, as well as our measure of marginal cost. Finally, we present our results, drawing heavily from Eichenbaum and Fisher (2003).

Our main findings can be summarized as follows. First, using postwar U.S. time-series data, we find strong evidence against the standard Calvo model ($\tau = 0$). This is true regardless of whether we allow for a structural break in monetary policy in the early 1980s, which a number of researchers argue occurred with the onset of the Volker–Greenspan era. Second, once we allow for a lag between the time that firms reoptimize and the time that they implement their new plans ($\tau > 0$), the model is no longer rejected. Third, allowing for measurement error in inflation also overturns the rejection of the standard Calvo model ($\tau = 0$). For reasons we discuss below, we are more comfortable with the second of the two resolutions.

Consider first the possibility that τ exceeds zero. If we use the full post-1959 sample period, we require that $\tau = 2$, that is, firms set prices two quarters in advance, to avoid rejecting the model. Frankly, we are skeptical that there is a six-month delay between when firms reoptimize their price plans and when they actually implement the new plan. So we do not view this as a plausible way of overturning the evidence against the standard Calvo model. Fortunately, once we allow for a break in monetary policy, the required value of τ arguably drops to one, which seems more reasonable on a priori grounds.

Turning to the other resolution, we find that even with independently and identically distributed (iid) classical measurement error, there is only marginal evidence against the standard Calvo model using the whole sample period. Once we allow for a break in monetary policy, we find virtually no evidence against the model. In addition, we cannot reject the null hypothesis that firms reoptimize prices, on average, once a year. This seems reasonable in light of the assumptions usually made in the literature. Of course a key question is: How large is the measurement error required to overturn the rejection of the model? We quantify the size of the measurement error using a variety of statistics. Our own view is that for the gross domestic product (GDP), Consumer Price Index (CPI), and personal consumption expenditures (PCE) deflator-based measures of inflation, the size of the required measurement error is reasonable, according to a variety of metrics documented in the article. We tentatively conclude that there is little evidence against the restrictions implied by the Calvo sticky price model.

The Calvo model of sticky prices

As discussed in the introduction, there are a variety of ways to model nominal rigidities in goods prices. Here we discuss the model of price setting associated with Calvo (1983). Since our objective is to derive the testable implications of this model per se, we do not embed it within a general equilibrium framework.

At time t, a final good, Y_t, is produced by a perfectly competitive firm. It does so by combining a continuum of intermediate goods, indexed by $j \in [0,1]$ using a constant returns to scale technology. We let P_t and P_{jt} denote the time t price of the final and intermediate good j, respectively. Profit maximization implies that the demand for intermediate good j is a decreasing function of the relative price of that good and an increasing function of aggregate output, Y_t.

The intermediate good $j \in [0,1]$ is produced by a monopolist that uses the following technology:

1) $Y_{jt} = A_t k_{jt}^{\alpha} L_{jt}^{1-\alpha}$,

where $0 < \alpha < 1$. Here, L_{jt} and k_{jt} denote time t labor and capital services used to produce the jth intermediate good, respectively. Intermediate firms rent capital and labor in perfectly competitive factor markets. The variable A_t denotes possible stochastic disturbances to technology.

Profits are distributed to the firms' owners at the end of each period. Let s_t denote the representative firm's real marginal cost, that is, the change in an optimizing firm's real total cost associated with increasing output by one unit.[5] Given our assumptions, marginal costs depend on the parameter α and factor prices, which the firm takes as given. The firm's time t profits are:

$$\left[\frac{P_{jt}}{P_t} - s_t\right] P_t Y_{jt},$$

where P_{jt} is firm j's price.

We assume that firms set prices according to a variant of the mechanism spelled out in Calvo (1983). In each period, a firm faces a constant probability, $1 - \theta$, of being able to reoptimize its nominal price. So, on average, a firm reoptimizes its price every $(1 - \theta)^{-1}$ periods. For example, if a period is one quarter and θ is 0.75, the firm reoptimizes on average once a year. We assume for simplicity that the firm's ability to reoptimize its price is independent across firms and time. For now we leave open the issue of what information set the firm has when it resets its price.

A standard assumption in the literature is that if the firm does not reoptimize its price, it updates its price according to the rule:

2) $P_{jt} = \bar{\pi} P_{jt-1}$,

where $\bar{\pi}$ is the long-run average gross rate of inflation (see, for example, Erceg, Henderson, and Levin, 2000, and Yun, 1996).[6]

As in Christiano, Eichenbaum, and Evans (2001), we interpret the Calvo price-setting mechanism as capturing firms' response to various costs of changing prices. The basic idea is that in the presence of these costs, firms fully optimize prices only periodically and follow simple rules for changing their prices at other times.

Let $\tilde{P}_t$ denote the value of P_{jt} set by a firm that can reoptimize at time t. Our notation does not allow $\tilde{P}_t$ to depend on j. We do this because, in models like ours, all firms that can reoptimize their price at time t choose the same price (see Woodford, 1996, and Yun, 1996). The firm chooses $\tilde{P}_t$ to maximize the expected present value of profits. We suppose that the firm sets $\tilde{P}_t$ on the basis of the information that it has at time $t - \tau$. When $\tau = 0$, the firm sees the realization of all time t variables when resetting its price. We refer to this version of the model as the standard Calvo model. The assumption that $\tau > 0$ is similar in spirit to the model in Mankiw and Reiss (2002), who think of firms as having flexible prices but "sticky" information sets. Alternatively one can imagine that resetting prices is a costly time consuming event for managers, so that prices must be set τ periods in advance. Given our assumptions, if the firm can reset its prices every period, then it would set its price, $\tilde{P}_t$, equal to a markup over the expected marginal cost conditional on information at $t - \tau$.

Log linearizing the first-order condition of the firm around the relevant steady state values, we obtain:

3) $$\hat{\tilde{p}}_t = E_{t-\tau}\left[\hat{s}_t + \sum_{l=1}^{\infty}(\beta\theta)^l(\hat{s}_{t+l} - \hat{s}_{t+l-1}) + \sum_{l=1}^{\infty}(\beta\theta)^l \hat{\pi}_{t+l}\right].$$

Here, $E_{t-\tau}$ denotes the conditional expectations operator. For example, $E_{t-\tau}\hat{s}_t$ denotes agents' expectations of $\hat{s}_t$ conditional on the information that they have at time $t - \tau$. In addition, $\tilde{p}_t = \tilde{P}_t / P_t$, and a hat over a variable indicates the percent deviation from its steady state value.

As noted by Christiano, Eichenbaum, and Evans (2001), several features of equation 3 are worth emphasizing. First, if inflation is expected to be at its steady state level and real marginal costs are expected to remain constant after time t, then the firm sets $\hat{\tilde{p}}_t = E_{t-\tau}\hat{s}_t$. Second, suppose the firm expects real marginal costs to be higher in the future than at time t. Anticipating those future marginal costs, the firm sets $\hat{\tilde{p}}_t$ higher than $E_{t-\tau}\hat{s}_t$. It does so because it understands that it may not be able to raise its price when those higher marginal costs materialize. Third, suppose firms expect inflation in the future to exceed its steady state level. To avoid a decline in its relative price, the firm incorporates expected future changes in the inflation rate into $\hat{\tilde{p}}_t$.

It follows from well-known results in the literature that the aggregate price level can be expressed as:[7]

4) $$P_t = \left[(1-\theta)\left(\tilde{P}_t\right)^{\frac{1}{1-\lambda}} + \theta\left(\bar{\pi}P_{t-1}\right)^{\frac{1}{1-\lambda}}\right]^{\frac{1}{1-\lambda}},$$

where $\lambda \in [1,\infty)$ is a parameter that controls the degree of substitutability of intermediate goods in the production of the final good. Log linearizing this relation and using it in conjunction with equation 3 implies that inflation satisfies:[8]

$$5)\quad \theta\hat{\pi}_t = \beta\theta E_{t-\tau}\hat{\pi}_{t+1} + (1-\beta\theta)(1-\theta)E_{t-\tau}\hat{s}_t.$$

While equation 5 is the focus of our empirical analysis, it is useful to note that it implies:

$$6)\quad \hat{\pi}_t = \frac{(1-\beta\theta)(1-\theta)}{\theta} E_{t-\tau}\sum_{j=0}^{\infty}\beta^j \hat{s}_{t+j}.$$

Relation 6 makes clear a central prediction of the model: Deviations of inflation from its steady state depend only on firms' expectations of current and future deviations of real marginal cost from its steady state value. So for example, in the short run, the growth rate of money, interest rates, or technology shocks affects inflation only by its effect on real marginal costs. In the long run, the rate of inflation depends on the average growth rate of money.

Assessing the empirical plausibility of the model

Here, we discuss the limited information strategy for testing the Calvo sticky price model pursued by Gali and Gertler (1999) and Eichenbaum and Fisher (2003), among others. The basic idea is to focus on the testable restrictions of the Calvo pricing model, while leaving unspecified other aspects of the economy.

To derive the testable implications of the Calvo model, it is convenient to define the random variable.

$$\psi_{t+1} = \left[\theta\hat{\pi}_t - \beta\theta\hat{\pi}_{t+1} - (1-\beta\theta)(1-\theta)\hat{s}_t\right].$$

Note that agents that reoptimize their price do so on the basis of their time $t-\tau$ information. The other prices that affect the time t inflation rate were already set on the basis of information before time $t-\tau$. This means that inflation is predetermined at time $t-\tau$. In principle, there are a variety of ways to test this assumption. For example, we could test whether any variable dated between time $t-\tau$ and t has explanatory power for time t inflation.

Here, we test this implication indirectly. Since $\hat{\pi}_t$ is in agents' time $t-\tau$ information set, equation 5 can be written as:

$$7)\quad E_{t-\tau}\psi_{t+1} = 0.$$

Relation 7 implies that the error agents make in forecasting the value of ψ_{t+1} when they reoptimize prices at time $t-\tau$ is uncorrelated with the information that they have at their disposal. Suppose that the $k \times 1$ vector of variables $X_{t-\tau}$ is in agents' time $t-\tau$ information set. Below, we refer to these variables as *instruments*. Then relation 7 implies the system of k equations:

$$8)\quad E_{t-\tau}\psi_{t+1}X_{t-\tau} = 0.$$

This in turn implies that the unconditional covariance between ψ_{t+1} and $X_{t-\tau}$ is equal to zero:

$$9)\quad E\psi_{t+1}X_{t-\tau} = 0.$$

Relation 9 provides us with a way to estimate the parameters of the model. Moreover, if the dimension of $X_{t-\tau}$ is greater than the number of parameters to be estimated, we can use these restrictions to test the model. To discuss our procedures for doing this, it is useful to recognize the dependence of ψ_{t+1} on the unknown value of (θ,β) by writing equation 9 as

$$10)\quad E\,[\psi_{t+1}(\theta,\beta)\,X_{t-\tau}] = 0.$$

Hansen (1982) provides conditions under which equation 10 can be used to consistently and efficiently estimate (θ,β) using generalized method of moments (GMM).[9] To discuss his procedure in our context, we define the vector

$$g_T(\theta,\beta) = \left(\frac{1}{T}\right)\sum_{t=1}^{T}\left[\psi_{t+1}(\theta,\beta)X_{t-\tau}\right].$$

Here T denotes the size of our sample. We also denote the true value of (θ,β) by (θ_0,β_0). The vector $g_T(\theta,\beta)$ is a consistent estimator of $E\,[\psi_{t+1}(\theta,\beta)\,X_{t-\tau}]$. The value of $E\,[\psi_{t+1}(\theta,\beta)\,X_{t-\tau}]$ is in general not equal to zero except at (θ_0,β_0). We estimate the parameter vector (θ_0,β_0) by choosing (θ,β) to make $g_T(\theta,\beta)$ as close to zero as possible in the sense of minimizing

$$11)\quad J_T = g_T(\theta,\beta)'W_T g_T(\theta,\beta).$$

Here, W_T is a symmetric positive definite matrix that can depend on sample information. Also the prime symbol (′) denotes the transpose operator. A given choice of W_T implies that we are choosing (θ,β) to minimize the sum of squares of k linear combinations of the elements of $g_T(\theta,\beta)$.

Hansen (1982) shows that the choice of W_T that minimizes the asymptotic covariance matrix of our estimator depends on the serial correlation properties

of the error term $\psi_{t+1}(\theta,\beta)$. In Eichenbaum and Fisher (2003), we show that the exact serial correlation properties of this error term depend on the value of τ. For example, if $\tau = 0$, then our model implies that $\psi_{t+1}(\theta,\beta)$ is serially uncorrelated. For $\tau \geq 1$, then $\psi_{t+1}(\theta,\beta)$ has a moving average representation of order 1. One does not have to impose this restriction in constructing an $\tau - 1$ estimate of W_T.[10] However, as we describe below, whether one does so has an important impact, in practice, on inference.

Hansen proves that the minimized value of the GMM criterion function, J_T, is asymptotically distributed as a χ^2 random variable with degrees of freedom equal to the difference between the number of unconditional moment restrictions imposed (k) and the number of parameters being estimated. We use this fact to test the restrictions imposed by the Calvo model.

Allowing for measurement error

The previous discussion assumes that inflation and real marginal costs are measured without error. We conclude this section by reviewing the results in Eichenbaum and Fisher (2003) about how measurement error affects the analysis. The possibility of measurement error in inflation is of particular interest to us. This is because a number of authors have noted that when they include a lagged inflation term in objects like ψ_{t+1}, it enters with a significant coefficient (see, for example, Gali and Gertler, 1999, and Fuhrer and Moore, 1995). These authors have interpreted this lagged term as evidence of firms that do not have rational expectations. Measurement error can provide an alternative interpretation of these findings.

There are well-known problems involved in measuring inflation. For example, it is widely believed that official CPI-based measures of inflation are biased due to changes in product quality and the benchmark basket of goods over time (see Shapiro and Wilcox, 1996). These problems are particularly severe when measuring rates of inflation over long periods. To the extent that this bias is constant, it does not affect our analysis. However, we must modify our econometric procedures to allow for time varying measurement error. Here we discuss the implications of classical measurement error.

Suppose that the econometrician has a measure of inflation π_t^m that is related to true inflation (π_t) via the relationship

$$\pi_t^m = \pi_t + u_t.$$

We suppose that u_t has a moving average representation of order q, denoted $MA(q)$, and that u_t is uncorrelated with π_t and the other variables in agents' information set at all leads and lags. This latter assumption defines what we mean when we say that u_t is classical measurement error. We continue to assume that agents see actual inflation.

To see how these assumptions impact our econometric procedures, consider the case in which τ is equal to zero and u_t is iid ($q = 0$). The econometrician now sees ϕ_{t+1}, which is a "polluted" version of the error term, ψ_{t+1}, that is the basis of the estimation procedure. The random variables ϕ_{t+1} and ψ_{t+1} are related as follows:

$$\phi_{t+1} = \psi_{t+1} + \theta\varepsilon_t - \theta\beta\varepsilon_{t+1}.$$

While ψ_{t+1} is uncorrelated with the elements of agents' time t information set, ϕ_{t+1} is correlated with π_t^m. Accordingly, measured time t inflation is not a valid instrument, that is, it cannot be included in X_t.

The presence of iid measurement error in inflation also means that ϕ_{t+1} has an $MA(1)$ representation. This affects the nature of the restrictions that the model imposes on the weighting matrix W_T. In Eichenbaum and Fisher (2003), we show how to estimate the volatility of the error term relative to the volatility of ϕ_{t+1}, as well as the contribution of measurement error to the volatility of measured inflation. This provides us with two metrics for assessing the size of the measurement error.

We refer the reader to Eichenbaum and Fisher (2003) for a discussion of the more general case in which u_t has a higher order MA representation. For our purposes, the key result is that when u_t has an $MA(q)$ structure, then ϕ_{t+1} has an $MA(q + 1)$ representation so that one must exclude $q + 1$ lags of inflation from the list of instruments, X_t. This structure also affects the restrictions that we can impose on the weighting matrix W_T.

We conclude this section by considering the possibility that real marginal costs are measured with a classical measurement error term that has an $MA(q)$ representation. If $\tau = 0$, then ϕ_{t+1} will have an $MA(q)$ representation, and one must exclude q lags of real marginal costs from the list of instruments, X_t. Below, we abstract from this source of measurement error and refer the reader to Eichenbaum and Fisher (2003) for an analysis of this case.

Measuring inflation and real marginal cost

Next, we discuss the measures of inflation and real marginal costs that we can use in our empirical analysis.

Inflation

Many different measures of inflation are of interest to economists and policymakers. Given the abstract nature of the Calvo model, there is no obviously *right* measure to use in our empirical analysis. In light of this, we considered four measures of inflation based on four measures of the aggregate price level: 1) the GDP deflator, 2) the price deflator for the nonfarm business sector (NFB), 3) the Consumer Price Index (CPI), and 4) the price deflator for personal consumption expenditures (PCE).[11] For each price measure, we constructed a measure of quarterly inflation over the period 1959–2001. In our empirical work, we measure $\hat{\pi}_T$ as the difference between actual time t inflation and the sample average of inflation.

Our different inflation measures are displayed in figure 1. As we can see, they behave in a similar manner over long periods of time. Inflation was low in the decade of the 1960s, then began a rapid rise with one peak in the early 1970s and another in the late 1970s. Thereafter, the different measures begin a long decline to very low levels by 2001. However, there are important differences between them over shorter periods. Since the Calvo model purports to account for movements in inflation over short periods of time, it is important to assess the robustness of our results using the different measures of inflation.

Real marginal costs

In our model, real marginal costs are given by the real product wage divided by the marginal product of labor. Given the production function we assumed in equation 1, this implies that real marginal cost is proportional to labor's share in national income, $W_t L_t / (P_t Y_t)$, where W_t is the nominal wage. In practice, we measure $W_t L_t$ as nominal labor compensation in the nonfarm business sector and we measure $P_t Y_t$ as nominal output of the nonfarm business sector. The variable $\hat{s}_t$ is then measured as the difference between the log of our measure of labor's share and its mean. This is a standard measure of $\hat{s}_t$, which has been used by Gali and Gertler (1999) and Sbordone (2001).

Rotemberg and Woodford (1999) discuss possible corrections to this measure that are appropriate for different assumptions about technology. These include corrections to take into account a non-constant elasticity of factor substitution between capital and labor and the presence of overhead costs and labor adjustment costs. In Eichenbaum and Fisher (2003), we discuss results for these alternative measures of marginal costs. In addition we allow for the possibility that firms require working capital to finance payments to variable factors of production. We argue in that paper that these corrections do not affect the qualitative nature of the results discussed below.

Panel A of figure 2 displays the log of our measure of real marginal cost and inflation measured using the GDP deflator for the sample 1959–2001. Notice that from the mid-1960s on, these two time series co-move positively. The bottom panel of figure 2 displays the dynamic correlations between real marginal cost at date t and inflation at date $t - k$, $k = -4, -3, \ldots, 4$. Clearly, inflation is positively correlated with past and future marginal costs.

Empirical results

Now, we present our empirical results.

The standard Calvo model

We begin by analyzing results based on the standard Calvo model, by which we mean the model described above with $\tau = 0$. In addition, we initially abstract from measurement error in inflation. We consider two specifications of the instrument vector X_t. Let Z_t denote the six dimensional vector consisting of the time t value of real marginal cost, quadratically detrended real GDP, inflation, the growth rate of an index of commodity prices, the spread between the annual interest rate on the ten-year Treasury bond and three-month Treasury bill, and the growth rate of nominal wages in the nonfarm business sector. This corresponds to the basic set of instruments used in Gali and Gertler (1999). In our first specification, X_t is given by

$$X_t^1 = \{1, Z_{t-j}, j = 0,1,2,3\}'.$$

As we discuss below, there are reasons to think that such a large set of instruments leads to misleading inference about the plausibility of the overidentifying restrictions implied by the model. With this in mind, we consider a second set of instruments given by

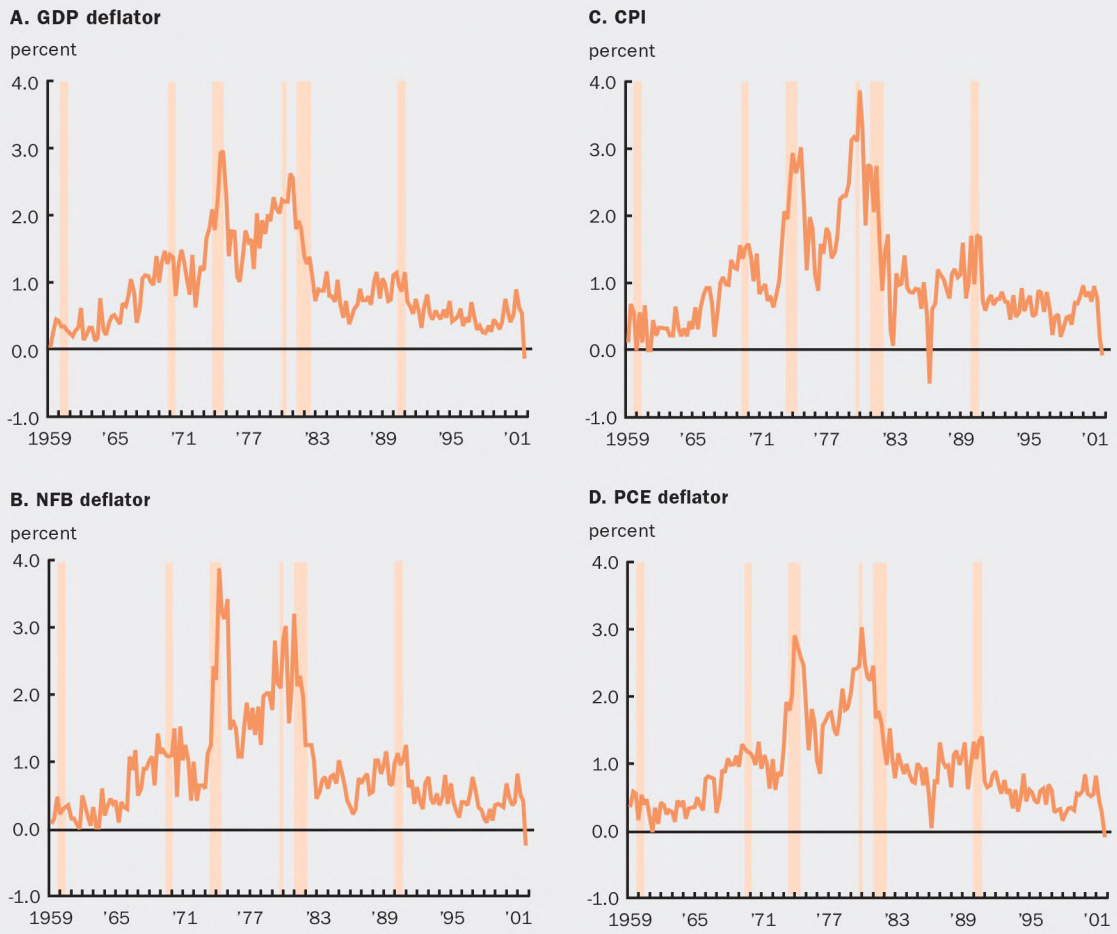
$$X_t^2 = \{1, Z_t, \psi_t\}'.$$

In Eichenbaum and Fisher (2003), we report that it is difficult to estimate β with great precision across the different specifications considered. Here, we summarize the results based on the assumption that $\beta = 0.99$.

Panel A of table 1, based on Eichenbaum and Fisher (2003), summarizes results when the standard Calvo model is estimated using the instrument vector X_t^1. We report our estimates of the parameter θ (standard

FIGURE 1

Four measures of inflation



Note: Shaded areas indicate official National Bureau of Economic Research recession periods.
Source: Authors' calculations based upon data from Haver Analytics.

error in parentheses) and the J_T statistic (p-value in brackets). The label L refers to the maximal degree of serial correlation that we allow for when estimating the weighting matrix W_T. We consider two values for L: 1) $L = 0$, which corresponds to the degree of serial correlation in ψ_{t+1} implied by this model, and 2) $L = 12$, the value used by Gali and Gertler (1999). Both values of L are admissible. But, by setting L to zero, we are imposing all of the restrictions implied by the model. This may lead to greater efficiency of our estimator and more power in our test of the overidentifying restrictions.

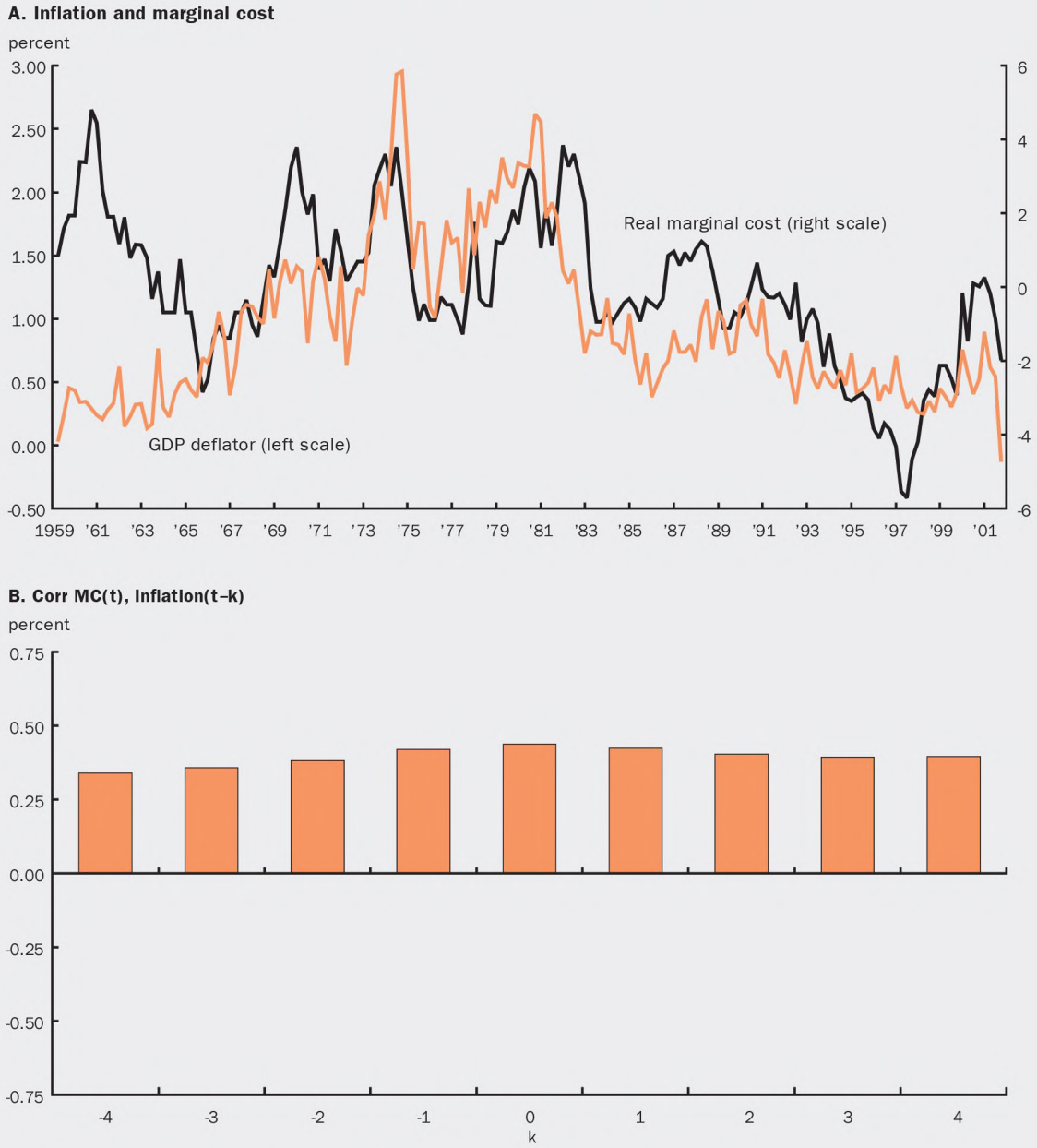
From table 1 we see that the parameter θ is estimated with relatively small standard errors. In addition, the point estimate itself is reasonably robust across the different inflation measures and the two values of L. The point estimates range from a low of 0.84 to a high of 0.91. This implies that, on average, firms wait between six and 11 quarters before reoptimizing their prices.

We hesitate to attribute too much importance to these point estimates. It is true that when $L = 12$ there is virtually no evidence against the model, at least based on the J_T statistic. This is consistent with results from Gali and Gertler (1999). However, when we set $L = 0$, the model is strongly rejected for three of the four inflation measures. In particular, the p-values for the non-CPI based inflation measures are well below 1 percent. Even in the CPI case, the p-value is 2 percent. Evidently, imposing all of the relevant restrictions implied by the model on the weighting matrix has an important impact on inference.

Panel B reports results based on the instrument vector X_t^2. A number of results are worth noting. First, our point estimates of θ are similar to those in panel A. Second, comparing the J_T statistics for $L = 12$ across

FIGURE 2

GDP deflator inflation and real marginal cost



Source: Authors' calculations based upon data from Haver Analytics.

the two instrument vectors, we find that there is more evidence against the model with the smaller list of instruments. However, with the instrument list X_t^2, the model is still not rejected at conventional significance levels for any inflation measure. Third, the model is decisively rejected when we set $L = 0$. Regardless of which inflation measure we use, the p-value of the J_T statistics is virtually zero. In light of these results, in Eichenbaum and Fisher (2003), we stress the importance of working with the smaller instrument set and imposing all of the model relevant restrictions on the weighting matrix W_T. For the rest of this article, we confine ourselves to results generated in this way.

An important maintained assumption of the previous results is that it is appropriate to use the entire sample period to estimate the model. In fact numerous observers

TABLE 1

Estimates of the standard model, 1959:Q1–2001:Q4

Inflation measure	$L = 0$ θ	$L = 0$ J_T	$L = 12$ θ	$L = 12$ J_T
A. Instruments: $\{1, Z_t, \ldots, Z_{t-3}\}'$				
GDP deflator	0.89	49.4	0.91	13.2
	(0.03)	[0.001]	(0.02)	[0.95]
NFB deflator	0.86	41.1	0.86	12.8
	(0.03)	[0.01]	(0.02)	[0.96]
CPI	0.88	38.9	0.86	12.6
	(0.05)	[0.02]	(0.02)	[0.96]
PCE deflator	0.87	44.8	0.88	12.8
	(0.02)	[0.004]	(0.02)	[0.96]
B. Instruments: $\{1, Z_t, \psi_{t-1}\}'$				
GDP deflator	0.90	28.2	0.91	10.3
	(0.05)	[9e–5]	(0.03)	[0.11]
NFB deflator	0.84	30.6	0.85	8.8
	(0.03)	[3e–5]	(0.03)	[0.18]
CPI	0.88	30.1	0.87	10.1
	(0.06)	[4e–5]	(0.03)	[0.12]
PCE deflator	0.87	36.9	0.89	11.5
	(0.04)	[2e–6]	(0.03)	[0.07]

Notes: The J_T statistics are distributed as χ^2 random variables with six and 23 degrees of freedom in panels A and B, respectively. Standard errors are in parentheses. P-values are in brackets. GDP is gross domestic product; NFB is nonfarm business; CPI is Consumer Price Index; and PCE is personal consumption expenditures.

Source: Authors' calculations based upon data from Haver Analytics.

have argued that there was an important change in the nature of monetary policy with the advent of the Volker disinflation in the early 1980s. Moreover, it is often argued that the Fed's operating procedures were different in the early 1980s than in the post-1982 period. Accordingly, we reestimated the standard Calvo model over two distinct subsamples: 1959:Q1–79:Q2 and 1982:Q3–2001:Q4.

Table 2 reports the subsample results (here $L = 0$ and the instrument vector is X_t^2). For the first sample period, there is strong evidence against the model for at least two measures of inflation. In particular, the p-values of the J_T statistic obtained using the NFB and PCE deflators are virtually zero. There is somewhat less evidence against the model when we use the GDP and CPI deflator based measures of inflation. Here the p-values are 0.04 and 0.02, respectively. In these cases, the point estimate of θ is 0.84. Taking sampling uncertainty into account, we would not reject the null hypothesis that, on average, firms wait about a year before reoptimizing their prices.

Turning to the second subsample, we see that there is substantially less evidence against the model. Here, the p-values associated with the J_T statistics obtained using the NFB, CPI, and PCE deflators are 0.06, 0.22, and 0.10, respectively. The only case in which we can reject the model at the 1 percent level of significance is when we use the GDP deflator to measure inflation. Interestingly, our point estimates of θ for specifications that are not strongly rejected are similar across subsamples. Again, taking sampling uncertainty into account, we would not reject the null that, on average, firms wait about a year before reoptimizing their prices.

Alternative timing assumptions

We now consider the results of estimating the model assuming $\tau = 1$ or $\tau = 2$. For these cases, our instrument list is given by X_{t-1}^2 and X_{t-2}^2, respectively. Panels A and B of table 3, based on Eichenbaum and Fisher (2003), summarize results for these cases. We begin by considering the full sample results. Two results are worth noting here. First, the point estimates of θ are similar across the different values of τ considered, including $\tau = 0$ discussed above. Second, when $\tau = 1$, the model's overidentifying restrictions are decisively

TABLE 2

Subsample estimates of standard model

Inflation measure	1959:Q1–79:Q2 θ	1959:Q1–79:Q2 J_T	1982:Q3–2001:Q4 θ	1982:Q3–2001:Q4 J_T
GDP deflator	0.84	13.4	0.86	17.0
	(0.04)	[0.04]	(0.04)	[0.009]
NFB deflator	0.74	21.7	0.86	12.2
	(0.03)	[0.001]	(0.04)	[0.06]
CPI	0.84	14.8	0.86	8.21
	(0.06)	[0.02]	(0.05)	[0.22]
PCE deflator	0.83	22.6	0.85	10.4
	(0.03)	[9e–4]	(0.04)	[0.10]

Notes: The J_T statistics are distributed as χ^2 random variables with six degrees of freedom. Standard errors are in parentheses. P-values are in brackets. GDP is gross domestic product; NFB is nonfarm business; CPI is Consumer Price Index; and PCE is personal consumption expenditures.

Source: Authors' calculations based upon data from Haver Analytics.

TABLE 3

Alternative timing assumptions

Inflation measure	Full sample θ	Full sample J_T	1959:Q1–79:Q2 θ	1959:Q1–79:Q2 J_T	1982:Q3–2001:Q4 θ	1982:Q3–2001:Q4 J_T
A. Prices chosen one period in advance						
GDP deflator	0.87	22.8	0.78	13.3	0.82	15.1
	(0.03)	[8e–4]	(0.04)	[0.04]	(0.02)	[0.02]
NFB deflator	0.82	28.0	0.66	22.1	0.81	12.7
	(0.04)	[9e–5]	(0.03)	[0.001]	(0.03)	[0.04]
CPI	0.86	18.4	0.72	15.7	0.83	6.37
	(0.04)	[0.005]	(0.03)	[0.02]	(0.04)	[0.38]
PCE deflator	0.83	22.9	0.75	12.9	0.82	8.51
	(0.02)	[8e–4]	(0.03)	[0.04]	(0.03)	[0.20]
B. Prices chosen two periods in advance						
GDP deflator	0.90	9.46	0.91	4.27	0.86	7.63
	(0.05)	[0.15]	(0.04)	[0.64]	(0.04)	[0.27]
NFB deflator	0.87	3.20	0.78	6.07	0.86	7.35
	(0.06)	[0.78]	(0.06)	[0.42]	(0.04)	[0.29]
CPI	0.86	10.8	0.81	5.02	0.85	4.08
	(0.05)	[0.09]	(0.04)	[0.54]	(0.05)	[0.67]
PCE deflator	0.88	7.72	0.84	5.28	0.88	3.52
	(0.04)	[0.26]	(0.03)	[0.51]	(0.06)	[0.74]

Notes: The J_T statistics are distributed as χ^2 random variables with six degrees of freedom. Standard errors are in parentheses. P-values are in brackets. GDP is gross domestic product; NFB is nonfarm business; CPI is Consumer Price Index; and PCE is personal consumption expenditures.
Source: Authors' calculations based upon data from Haver Analytics.

rejected. The p-value associated with the J_T statistic corresponding to every measure of inflation is very small. However, when $\tau = 2$, there is very little evidence against the model. In no case is the p-value less than 0.08. Our view is that this is somewhat of a Pyrrhic victory for the Calvo model. It is entirely possible that there is some delay between when firms reoptimize their price plans and when they actually implement the new plan. But, it is not clear that a six-month delay is a credible assumption.

Consider next the results obtained over the sample period 1959:Q1–72:Q2. A number of interesting findings emerge. First, when $\tau = 1$ the point estimates of θ are substantially smaller than the corresponding estimates obtained over the full sample. For example, with CPI inflation, the point estimate of θ falls from 0.86 to 0.72. Second, with the exception of NFB inflation, there is only marginal evidence against the model when $\tau = 1$. Third, there is virtually no evidence against the model when $\tau = 2$.

Finally, consider the results obtained over the sample period 1982:Q3–2001:Q4. Notice that the point estimates of θ are larger than the corresponding estimates obtained for the first subsample for all values of τ. Perhaps more importantly, there is relatively little evidence against the model with $\tau = 1$ and virtually no evidence against the model when $\tau = 2$.

Viewed as a whole, these results indicate that the Calvo model performs reasonably well if we allow for a split in the sample period and for a lag of roughly one quarter between when firms reoptimize their price plan and when they actually implement the new plan.

Impact of measurement error in inflation

We now consider the results of estimating the model allowing for the possibility that inflation is measured with error of the form

$$12)\quad u_t = \varepsilon_t + \gamma_1\varepsilon_{t-1} + \gamma_2\varepsilon_{t-2} + \ldots + \gamma_q\varepsilon_{t-q}.$$

The model is otherwise the standard Calvo model ($\tau = 0$). For each measure of inflation, we report results for the minimal level of q, such that the overidentifying restrictions of the model are not rejected at the 1 percent level.

To motivate why this model of the measurement could improve the model's performance, figure 3 displays the basic Euler equation errors emerging from the standard Calvo model estimated over the full sample period. These errors are negatively serially correlated

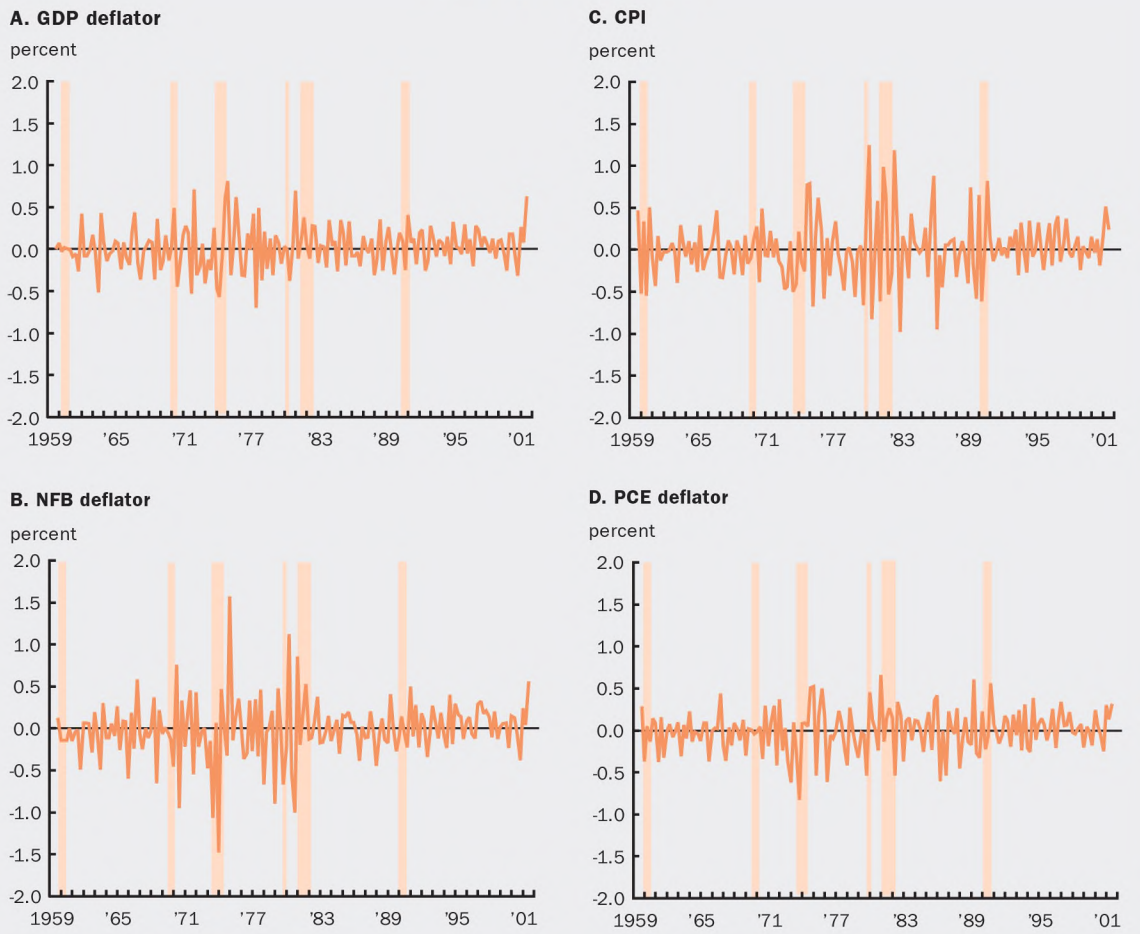
(in all cases the first order correlation coefficient is about –0.25) and are plausibly modeled with a low-order moving average representation. As relation 12 indicates, even iid measurement error can generate a time series for ϕ_t that is negatively serially correlated.

Panel A of table 4 reports the results of estimating the model, allowing for measurement error, based on the full sample period. To help assess the magnitude of the measurement error, the column labeled Γ_1 reports our estimate of the ratio of the variance of true inflation to the variance of the measurement error. This is one measure of the extent of measurement error in the inflation data. Below we refer to Γ_1 as the signal to noise ratio in the inflation data. The column labeled Γ_2 reports our estimate of the percentage of the variance of the composite error term ϕ_t due to classical measurement error that is observed by the econometrician.

A number of key results are worth noting here. First, for all measures of inflation, allowing for iid measurement error overturns the strong rejection of the standard model reported in table 1. Indeed, for the GDP and NFB measures, the overidentifying restrictions cannot be rejected at even the 10 percent level. For the PCE deflator, these restrictions are not rejected at the 4 percent significance level. Second, taking sampling uncertainty into account, our point estimates of the parameter θ are reasonably similar to those reported in table 1. Third, measurement error appears to be more important for the NFB deflator. For the GDP, CPI, and PCE deflators, the ratio of the variance of true inflation to the variance of the measurement error (Γ_1) is 18.4, 13.7, and 19.5, respectively. Evidently, the signal to noise ratio in these inflation measures is high. In the case of the NFB deflator, this ratio is

FIGURE 3

Euler equation residuals



Note: Shaded areas indicate official National Bureau of Economic Research recession periods.
Source: Authors' calculations based upon data from Haver Analytics.

roughly 7.43, so the signal to noise ratio is lower. The percentage of the variance of the composite error term observed by the econometrician due to classical measurement error (Γ_2) is 0.48, 0.45, and 0.46, for the GDP, CPI, and PCE deflators, respectively. But for the NFB deflator, this ratio is roughly 0.73. So based on either the Γ_1 or the Γ_2 statistic, there appears to be more noise associated with the NFB deflator.

Panels B and C in table 4 report our subsample results. Note that for every measure of inflation, there is virtually no evidence against the model in either sample period, once we allow for even iid measurement error. Our point estimates of θ are higher in the second sample period, implausibly so for the NFB deflator. But taking sampling uncertainty into account, one cannot reject the hypothesis, for any measure of inflation or in either subsample period, that firms reoptimize prices, on average, once a year ($\theta = 0.75$).

Turning to our measures of the importance of classical measurement error, a number of results are worth noting. First, in the pre-1979 sample period, the importance of measurement error, assessed using either the Γ_1 or Γ_2 statistic, is highest for the NFB measure of inflation. Indeed, the value of the Γ_2 statistic is so high (0.80) that we are led to conclude that either 1) the NFB is a relatively unreliable measure of true inflation in the first period, or 2) our model of measurement error is implausible. Second, in the post-1982 sample period, NFB inflation has estimated measurement error properties that are quite similar to those of the GDP and PCE deflators. Third, there is a substantial decline in the signal to noise ratio for all three of the inflation measures in the second subsample period.

Viewed as a whole, these results indicate that allowing for classical measurement error results is a large improvement in the model's performance.

TABLE 4

Measurement error in inflation

Inflation measure	θ	J_T	Γ_1	Γ_2
A. Full sample				
GDP deflator	0.91 (0.04)	10.3 [0.11]	18.4	0.48
NFB deflator	0.90 (0.05)	9.54 [0.15]	7.43	0.73
CPI	0.90 (0.06)	16.6 [0.01]	13.7	0.45
PCE deflator	0.91 (0.04)	12.9 [0.04]	19.5	0.46
B. 1959:Q1–79:Q2				
GDP deflator	0.86 (0.04)	4.97 [0.55]	14.8	0.48
NFB deflator	0.82 (0.05)	6.42 [0.38]	5.64	0.80
CPI	0.85 (0.06)	5.87 [0.44]	24.1	0.38
PCE deflator	0.87 (0.05)	5.94 [0.43]	31.8	0.32
C. 1982:Q3–2001:Q4				
GDP deflator	0.92 (0.07)	6.39 [0.38]	3.10	0.59
NFB deflator	0.93 (0.08)	6.14 [0.41]	4.27	0.47
CPI	0.88 (0.06)	5.48 [0.48]	2.12	0.59
PCE deflator	0.92 (0.09)	4.47 [0.61]	3.07	0.65

Note: Γ_1 is the ratio of the variance of the true inflation rate to the variance of the measurement error component; Γ_2 is the fraction of the variance of ϕ_{t+1} due to measurement error. The J_T statistics are distributed as χ^2 random variables with six degrees of freedom. Standard errors are in parentheses. P-values are in brackets. GDP is gross domestic product; NFB is nonfarm business; CPI is Consumer Price Index; and PCE is personal consumption expenditures.
Source: Authors' calculations based upon data from Haver Analytics.

Conclusion

This article discussed the empirical performance of the Calvo model of sticky goods prices. We argued there is overwhelming evidence against this model. But this evidence was generated under three key maintained assumptions. First, there is no lag between the time firms reoptimize their price plans and the time they implement those plans. Second, there is no measurement error in inflation. Finally, monetary policy was the same in the pre-1979 period and the post-1982 period.

Drawing heavily from results in Eichenbaum and Fisher (2003), we discussed the impact of relaxing each of these assumptions. Relaxing the first and third assumptions overturns the evidence against the model, if we are willing to assume that firms wait roughly one quarter before implementing new price plans. Relaxing just the second assumption by allowing for iid classical measurement error is sufficient by itself to render the evidence against the standard Calvo model marginal. If we relax both the second and third assumptions, we find virtually no evidence against the model. Moreover, we find little evidence against the view that firms reoptimize their prices, on average, once a year.

NOTES

[1]See for example, Chari, Kehoe, and McGrattan (2000), Christiano, Eichenbaum, and Evans (2001), Erceg, Henderson, and Levin (2000), Gali and Gertler (1999), Rotemberg and Woodford (1997), and Yun (1996).

[2]For example, Burstein (2002) shows that for moderate changes in the growth rate of money (less than or equal to 5 percent on a quarterly basis), traditional time-dependent models are a good approximation of state-dependent models.

[3]See, for example, Ireland (1997) and Cho and Moreno (2002).

[4]See, for example, Christiano, Eichenbaum, and Evans (2001) and Altig, Christiano, Eichenbaum, and Linde (2003), respectively.

[5]We do not index s_t by j, because all firms have identical marginal costs.

[6]Others, like Dotsey, King, and Wolman (1999), and Woodford (1996), assume $P_{jt} = P_{jt-1}$. Christiano, Eichenbaum, and Evans (2001) also consider a dynamic indexing scheme in which $P_{jt} = \pi_{t-1} P_{jt-1}$. In Eichenbaum and Fisher (2003), we evaluate the performance of the Calvo model under these alternative specifications.

[7]See, for example, Calvo (1983).

[8]For a proof of this, see Woodford (1996) or Yun (1996).

[9]The key assumption is that $\{\hat{\pi}_t, \hat{s}_t, X_t\}$ is a stationary and ergodic process. We also require that $k \geq 2$.

[10]That is, when constructing an estimate of W_T, one could allow for higher order serial correlation in the error term than the theory implies.

[11]Detailed data sources are discussed in the appendix.

APPENDIX

Our data are from the Haver Analytics database. For each data series below, we provide a brief description and, in parentheses, the Haver codes for the series used.

- Price measures: GDP deflator is the ratio of nominal GDP (GDP) and real chain-weighted GDP (GDPH); nonfarm business deflator (LXNFI); Consumer Price Index (PCU); and personal consumption expenditures deflator (JCBM2).

- Real marginal costs: Share of labor income in nominal output for the nonfarm business sector, which is proportional to the U.S. Bureau of Labor Statistics' measure of nominal unit labor costs divided by the nonfarm business deflator (LXNFU/LXNFI).

- Instruments: Quadratically detrended real GDP is the residual of a linear regression of real GDP (GDPH) on a constant, t and t^2; inflation is the first difference of the log of the price measures; the index of commodity prices is the Commodity Research Bureau's index of prices of all commodities (PZALL); the interest rate spread is the difference between a composite of yields on interest rates on Treasury bonds of maturity ten years and greater (FLGT) and the interest rate on three-month Treasury bills (FTBS3); and growth rate of nominal wages is the first difference of the log of nominal compensation in the nonfarm business sector.

REFERENCES

Altig, David, Lawrence Christiano, Martin Eichenbaum, and Jesper Linde, 2003, "Monetary policy and the diffusion of technology shocks," Northwestern University, manuscript.

Burstein, A., 2002, "Inflation and output dynamics with state dependent pricing decisions," University of Michigan, manuscript.

Calvo, Guillermo, 1983, "Staggered prices in a utility-maximizing framework," *Journal of Monetary Economics*, Vol. 12, No. 3, pp. 383–398.

Chari, V. V., Patrick Kehoe, and Ellen McGrattan, 2000, "Sticky price models of the business cycle: Can the contract multiplier solve the persistence problem?," *Econometrica*, Vol. 68, No. 5, September, pp. 1151–1179.

Cho, S., and A. Moreno, 2002, "A structural estimation and interpretation of the new Keynesian macro model," Columbia University, manuscript.

Christiano, Lawrence, Martin, Eichenbaum, and Charles Evans, 2001, "Nominal rigidities and the dynamic effects of a shock to monetary policy," Northwestern University, manuscript.

Dotsey, Michael, Robert, G. King, and Alexander L. Wolman, 1999, "State-dependent pricing and the general equilibrium dynamics of money and output," *Quarterly Journal of Economics*, Vol. 114, No. 2, May, pp. 655–690.

Eichenbaum, M., and J. Fisher, 2003, "Testing models of sticky prices," Northwestern University, manuscript.

Erceg, Christopher, J., Dale W. Henderson, and Andrew T. Levin, 2000, "Optimal monetary policy with staggered wage and price contracts," *Journal of Monetary Economics*, Vol. 46, No. 2, October, pp. 281–313.

Fuhrer, Jeffrey, and Jeffery Moore, 1995, "Inflation persistence," *Quarterly Journal of Economics*, Vol. 110, No. 1, pp. 127–159.

Gali, Jordi, and Mark Gertler, 1999, "Inflation dynamics: A structural econometric analysis," *Journal of Monetary Economics*, Vol. 44, pp. 195–222.

Hansen, Lars P., 1982, "Large sample properties of generalized method of moments estimators," *Econometrica*, Vol. 50, pp. 1029–1054.

Hansen, Lars P., and Kenneth J. Singleton, 1982, "Generalized instrumental variables estimation of nonlinear rational expectations models," *Econometrica*, Vol. 50, pp. 1269–1286.

Ireland, Peter, 1997, "A small, structural, quarterly model for monetary policy evaluation," *Carnegie Rochester Conference Series on Public Policy*, Vol. 47, pp. 83–108.

Kim, Jinill, 2000, "Constructing and estimating a realistic optimizing model of monetary policy," *Journal of Monetary Economics*, Vol. 45, No. 2, April, pp. 329–359.

Mankiw, N. G., and R. Reiss, 2002, "Sticky information: A proposal to replace the new Keynesian Phillips Curve," *Quarterly Journal of Economics*, Vol. 117, November, pp. 1295–1328.

Rotemberg, Julio J., and Michael Woodford, 1999, "The cyclical behavior of prices and costs," *Handbook of Monetary Economics*, Vol. 1B, Michael Woodford and John Taylor (eds.), Amsterdam; New York, and Oxford: Elsevier Science, North-Holland.

_____, 1997, "An optimization-based econometric framework for the evaluation of monetary policy," *Macroeconomics Annual*, National Bureau of Economic Research, Vol. 12.

Sbordone, Argia, 2001, "An optimizing model of U.S. wage and price dynamics," Rutgers University, manuscript.

Shapiro, Matthew, and David W. Wilcox, 1996, "Mismeasurement in the Consumer Price Index: An evaluation," *Macroeconomics Annual*, National Bureau of Economic Research, Vol. 11, pp. 93–142.

Taylor, John, 1980, "Aggregate dynamics and staggered contracts," *Journal of Political Economy*, Vol. 88, February.

Woodford, Michael, 1996, "Control of the public debt: A requirement for price stability," National Bureau of Economic Research, working paper, No. 5684.

Yun, Tack, 1996, "Nominal price rigidity, money supply endogeneity, and business cycles," *Journal of Monetary Economics*, Vol. 37, No. 2, pp. 345–370.