International Journal of Health Care Finance and Economics, 5, 89–118, 2005

Who Really Pays for Health Insurance? The Incidence of Employer-Provided Health Insurance with Sticky Nominal Wages

BENJAMIN D. SOMMERS sommers@fas.harvard.edu

Harvard Ph.D. Program in Health Policy, 79 John F. Kennedy Street, Cambridge, MA 02138

This paper addresses two seeming paradoxes in the realm of employer-provided health insurance: First, businesses consistently claim that they bear the burden of the insurance they provide for employees, despite theory and empirical evidence indicating that workers bear the full incidence. Second, benefit generosity and the percentage of premiums paid by employers have decreased in recent decades, despite the preferential tax treatment of employer-paid benefits relative to wages—trends unexplained by the standard incidence model. This paper offers a revised incidence model based on nominal wage rigidity, in an attempt to explain these paradoxes. The model predicts that when the nominal wage constraint binds, some of the burden of increasing insurance premiums will fall on firms, particularly small companies with low-wage employees. In response, firms will reduce employment, decrease benefit generosity, and require larger employee premium contributions. Using Current Population Survey data from 2000–2001, I find evidence for this kind of wage rigidity and its associated impact on the employment and premium contributions of low-wage insured workers during a period of rapid premium growth.

Keywords: employer-provided health insurance, wage stickiness, premium contributions

JEL classification: I11, J32

Introduction

> "*As far as employers are concerned, we see no end in sight... Employers will be increasing not only cost sharing in the premium paid each month but in every possible way, with either a co-payment or co-insurance.*"
>
> —Helen Darling, former Benefits Manager for Xerox Corporation[1]

> *"Many small companies have no choice but to eat the premium increases."*
>
> —Bill Lindsay, CEO of Benefit Management & Design Inc.[2]

The standard economic view of employer-provided benefits is that firms only nominally pay for health insurance, pensions, and the like, while the actual costs are born by employees through lower wages. In short, the incidence falls fully on workers. Summers (1989) outlines the standard theory of employer-provided health insurance: Mandated benefits cause shifts in both the supply and demand of labor, leading to an equilibrium wage that absorbs most, if not all, of the cost of an insurance benefit. Empirical work on equalizing wage differentials for benefit mandates supports this prediction (Gruber and Krueger, 1991; Gruber, 1994).

Most non-economists, however, believe that firms bear the cost of these benefits. The traditional model fails to explain the claims by businesses, especially small ones, that they cannot "afford" health insurance for employees as premiums grow rapidly. Most of the attention this topic has received in the literature has taken the form of editorial commentary rather than a formal model (Krueger and Reinhardt, 1994; Pauly, 1997). Furthermore, the standard model does not explain the trends of growing employee contributions to insurance premiums and decreasing benefit generosity. These phenomena are difficult to reconcile with the preferential tax treatment received by employer-provided benefits, as any compensation shifted away from insurance towards wages increases costs to the firm without affecting labor supply. While several papers have examined these trends (e.g. Dranove, Spier and Baker, 2000; Gruber and McKnight, 2003), none has made an explicit connection between these apparent violations of the classical incidence model and the more general skepticism of non-economists regarding who really pays for health insurance.

This paper explores the following assumption about wage stickiness in an attempt to reconcile the views of economists and non-economists: Workers do not like to see their wages go down. In practice, this assumption means that *nominal* wage cuts are costly to firms. Two possible sources of this cost are increased shirking by disgruntled employees and increased turnover. However, this paper does not attempt to document why wages are sticky, or why other forms of compensation—such as the employee premium contribution—may not exhibit the same rigidity, although research indicates that perceptions of fairness play an important role (Kahneman, Knetsch and Thaler, 1986). Rather, the focus of the paper is exploring the implications of these constraints. Nonetheless, research on nominal wage stickiness offers support for the validity of this key assumption (Kahn, 1997; Card and Hyslop, 1997). The model in this paper considers a firm that pays its employees a combination of cash wages and health insurance. Over time, the cost of insurance increases, and the firm must adjust its compensation and output accordingly.

The basic structure of the paper is this: Section 1 presents a graphical and analytical version of the base case, a two-period production model in which the firm maximizes profit by setting wage and quantity. Section 2 adds a third variable to the firm's control, the employee premium contribution, and also considers the impact of the federal tax subsidy for employer insurance costs. Section 3 presents empirical tests of the model. Section 4 concludes.

1. The Basic Model

Before examining a formal model, consider a diagram (Figure 1) that summarizes the paper's main concepts. In Period 0, labor supply is upward-sloping for all values of the wage above the current wage, but workers refuse to supply any labor if their wage falls. In Period 1, the real cost of health insurance has gone up, leading to a dollar-for-dollar leftward shift in the labor demand curve. Labor supply shifts too, taking account of the increased value of insurance benefits, but retaining the workers' unwillingness to work for less than the

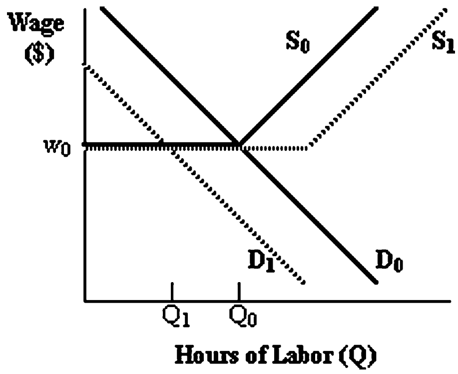

Figure 1. Employer-provided benefits with nominal wage rigidity.

Period 0 wage. The result of premium growth combined with nominal wage rigidity is that real wages remain artificially high, the quantity of labor decreases (unemployment rises), and firm profit falls. Analytically, these results parallel those of a binding minimum wage coupled with a new mandated benefit.

Three main factors may enable firms to avoid this outcome or at least mitigate the resulting loss in profit: First, if general inflation is sufficiently high, firms can recoup the added premium costs without decreasing nominal wages by allowing inflation to erode the real wage. Second, firms can reduce the generosity of the insurance benefit, by reducing the extent of covered services or shifting towards cheaper plan designs (such as from fee-for-service to managed care). Third, firms can require employees to contribute directly towards premiums. We turn to the formal model to examine these results.

Consider a profit-maximizing firm facing downward-sloping demand, $p(Q)$; I will return to this topic later to discuss which of the model's results hinge on the assumption of imperfectly competitive product markets. For simplicity, I assume that labor is the only input, and production exhibits constant returns to scale. Workers are paid with wages (w) and health insurance benefits (b), both expressed in terms of dollar per unit output. These parameters are normalizations of the more familiar metrics of dollars per hour and monthly premiums, respectively, which could be derived from w and b by multiplying by units of output per hour or per month. In Period 0, the firm is a price-taker for wages and benefits, offering the market's "standard" employment package.[3] This differs slightly from the labor supply in Figure 1, but the intuition is the same. The firm sets Q to solve: Max $\pi = p(Q_0) \cdot Q_0 - Q_0 \cdot (w_0 + b)$.

In Period 1, i denotes inflation and g denotes real growth in health insurance premiums. Note that g encompasses more than just excess medical inflation; it also includes new health care products that typically are more expensive than existing technologies. Thus, the nominal cost of insurance premiums in Period 1 is equal to $b \cdot g \cdot i$. For the time being, I assume that b is fixed in both periods—the firm cannot choose to drop coverage below the "standard" employment package or require an employee premium contribution—but this assumption will be relaxed in Section 2. In Period 1, demand $p(Q)$ and real productivity are unchanged. The firm now maximizes over two variables, quantity (Q_1) and wage (w_1);

w_1 is expressed in nominal terms, so it alone will not be multiplied by the inflation term i in the profit maximization problem. The firm now operates under two wage constraints:

WC1: Total real compensation is equal to or greater than it was in Period 0. This is the "competition constraint," maintaining the assumption that labor markets are competitive.
WC2: Nominal wages do not decrease. This is the "nominal wage constraint," representing the extreme case in which a nominal wage cut is infinitely costly to the firm.

As discussed in Summers (1989), the workers' valuation of the noncash benefit is an important parameter. For analytical tractability, I assume that workers value the benefits at their real cost to the firm in both periods (this will change when I consider the tax wedge in Section 2).

The firm faces this maximization problem in Period 1:

$$\begin{aligned} &\max_{Q_1, w_1} \quad \pi = i \cdot p(Q_1) \cdot Q_1 - Q_1 \cdot (w_1 + b \cdot g \cdot i) \\ &\text{s.t.} \quad w_1/i + b \cdot g \geq w_0 + b \quad \text{(WC1)} \\ &\text{s.t.} \quad w_1 \geq w_0 \quad \text{(WC2)} \end{aligned} \tag{1}$$

The case where only WC_1 binds is the textbook example in economic theory: The firm is able to shift the full cost of increasing insurance onto employees. Using WC1, we can easily solve:

$$w_1 = i(w_0 + b - b \cdot g) \tag{2}$$

If there is no real growth in the cost of insurance ($g = 1$), Eq. (2) simplifies to $w_1 = i \cdot w_0$. With no change in the relative costs of wages vs. insurance, the firm pays the same real wage, and quantity is unchanged. But real premium growth ($g > 1$) leads to a decrease in the real wage. The added cost of insurance gets fully shifted onto the employee through inflation-eroded wages, even as the nominal wage increases. The firm's marginal cost is unchanged, so quantity and real profit are unaffected by the increase in premiums. This is the standard economic prediction.

The model produces novel results only when the nominal wage constraint binds. Let $g = 1 + r_g$ and $i = 1 + r_i$, where r_g and r_i are the real rate of premium growth and the general inflation rate, respectively. Combining the two constraints yields the following condition for when the nominal wage constraint (WC2) will bind (see Appendix A for the derivation):

$$r_g \geq (w/b) \cdot (r_i/i) \qquad \text{Ineq. (1)}$$

When WC2 binds, the firm faces a problem that does not fit in the traditional model of employer-provided benefits. Increasing insurance costs and a rigid wage mean that the firm's marginal cost has gone up (for now, we consider the firm's response assuming it cannot reduce the generosity of benefits). With an increased marginal cost and downward-sloping demand, the effect of the wage constraint is clear: The firm will reduce quantity. This is simply standard price-setting behavior. The result corresponds to the common complaint from businesses that expensive benefits cut into their profits. When the nominal wage constraint binds, increasing health care costs reduce a firm's profit and lead it to increase prices. In addition, cutting quantity may allow firms to concentrate labor hours among fewer workers, a profitable move since insurance can be seen as a fixed per-worker expense (Cutler and Madrian, 1998). As a brief aside, if we consider a firm that faces a competitive product market, rather than the assumed downward-sloping demand, the prediction is even simpler—the firm simply eats the losses of premium growth and cannot raise prices. In either case, insurance benefits cut into firm profits.

The comparative statics for price, quantity, and profit for the basic model are straightforward—see Appendix A for their derivation. Firm profit is decreasing in the size of the insurance benefit (b) and the rate of premium growth (g), since both factors increase the added marginal cost to the firm. Profit is increasing in the inflation rate (i) and the initial wage (w_0), since larger values of these variables allow firms to shift more of the premium increase onto workers before hitting the nominal wage constraint. The degree of price and quantity distortion varies inversely with these changes in profit (i.e. factors that increase profits imply less price and quantity distortion).

1.1. Incidence

While these comparative statics suggest that firms facing rapidly increasing health care costs can experience decreasing profits as a result, workers are not insulated from the costs of premium growth. Workers still bear most of the burden of their insurance, in the form of inflation-eroded wages. Furthermore, this profit effect on the firms is accompanied by a decrease in quantity, and quantity here refers to both output and labor input. The incidence of growing insurance costs when the nominal wage constraint binds therefore falls on three groups: (1) The firm, through lost profits; (2) Period 1 workers, through inflation-eroded wages; and (3) Period 0 workers who are laid off in Period 1.[4]

Having outlined the effects of the wage constraint, we return to the question, "When will the nominal wage constraint bind?" Earlier in this section, we saw that WC2 would bind if:

$$r_g \geq (w/b) \cdot (r_i/i) \qquad \text{Ineq. (1)}$$

Note that Ineq. (1) does not depend on the functional form of the firm's cost, since the inequality was derived only from the nominal wage constraint. Appendix B further shows that this result does not depend on a competitive labor market. Thus, Ineq. (1) offers a generalizable description of when nominal wage rigidity will trigger the price/profit responses outlined above. Several predictions emerge. Most obvious is that, all else equal, firms facing rapid premium growth (r_g) are more likely to bear some of the burden of their workers' insurance. But more interesting are the implications of the right-hand side of the inequality: Anything that decreases the expression $(w/b) \cdot (r_i/i)$ makes it more likely that a firm bears some burden. Thus, the model offers several predictions as to which firms will be most affected by growing insurance costs:

First, firms facing low general inflation (r_i/i) will be more likely to bear some burden. The wage constraint forces firms to use inflation to erode the real wage, in order to shift increasing insurance costs onto workers. Firms in geographical regions or time periods with lower inflation will be less able to do so. This prediction offers a testable hypothesis explored in Section 3.

Second, firms whose workers receive a large share of their total compensation in the form of health insurance benefits will be more likely to bear some burden. Workers with high relative wages (w/b) have a significant buffer from general inflation that allows firms to shift the increasing health care costs. For example, a firm whose employees earn \$100,000 a year and receive benefits worth \$5,000 can easily cover even a 20% hike in premiums when inflation is 2% (the \$1000 of extra insurance cost is only half of each worker's "inflation raise" of \$2000).

Third, firm size may indeed be a factor in determining the degree to which rising premiums are a burden, despite economists' general dismissal of this claim. Though economists recognize that large firms face smaller loading fees for insurance, the standard model suggests that this should only affect wages, not profits. For instance, Fuchs (1994) points to law firms as evidence that there is no reason to think that small firms cannot provide health insurance to their workers. However, Ineq. (1) indicates that any firm with a low ratio w/b will be more likely to bear some burden of insurance costs. Given the economy of scale in purchasing insurance, small firms have a lower wage-benefit ratio than larger firms offering their workers the same level of health coverage and wages (since they pay a lower price (b). Taking these two effects together—the wage effect and the economy of scale in buying insurance—Ineq. (1) predicts that small firms with low-wage workers are most likely to bear the burden of premium growth. This jibes with the standard refrain from industry, and it also explains Fuchs's observation: Law practices do not struggle to purchase insurance because such firms satisfy only one of these two conditions—they are small businesses, but on average their employees are certainly not low-wage.[5]

Nationwide, there is a greater range in wages than in per-person insurance costs for workers. In 2001, annual income ranged from \$15,000 to \$100,000 among full-time working adults at the 10th and 95th percentiles, respectively (CPS, 2001). During the same year, the average annual cost of employer-provided insurance ranged from \$2400 for an individual HMO policy to \$7600 for fee-for-service family coverage (KFF, 2001). These numbers indicate that low-wage workers with bottom-barrel insurance have much lower w/b ratios—roughly 6 on average—than high-wage earners with generous insurance, whose ratios can exceed 20. Furthermore, worker heterogeneity in coverage types could lead to even wider disparities in the w/b ratio, as a low-wage worker with family HMO coverage would spend roughly \$6500 on coverage (a ratio of less than 2.5), while a high-wage worker with an individual fee-for-service plan would spend roughly \$2900 (a ratio of over 30) (*ibid*). Thus, depending on inflation and premium growth, we would expect small firms with low-wage workers to feel the pinch, while larger firms and small firms with high-wage workers would act just as the standard incidence model predicts.

Table 1 provides some numbers outlining this effect. For each pair of inflation rate and wage/benefit ratio, the table provides the threshold rate of premium growth needed for the nominal wage constraint to bind. The threshold values indicate that there are only some

Table 1. Threshold growth in insurance premiums for nominal wage constraint to bind.

	General inflation			
	1%	2%	3%	5%
Employee type				
Low wage ($20,000)				
Individual HMO ($2400)	8	16	24	40
Family HMO ($6500)	3	6	9	15
Middle wage ($50,000)				
Individual HMO ($2400)	21	41	61	99
Family HMO ($6500)	8	15	22	37
High wage ($100,000)				
Individual fee-for-service ($2900)	34	68	100	164
Family fee-for-service ($7600)	13	26	38	63

Table 2. Historical premium growth & inflation: 1991–2001.

Year	Premium growth (%)	Inflation (%)
1988	12.0	4.1
1989	18.0	4.8
1990	14.0	5.4
1991	12.0*	4.2
1992	10.0*	3.0
1993	8.5	3.0
1994	6.0*	2.6
1995	3.0*	2.8
1996	0.8	3.0
1997	2.0*	2.3
1998	4.0*	1.6
1999	4.8	2.2
2000	8.3	3.4
2001	11.0	2.8

Sources: Premium growth data from KFF (2001, 2002); inflation data from BLS (2002).
*Data extrapolated from KFF graphs, for missing years.

subgroups of workers for which the constraint is likely to bind with regularity. Given that health insurance premium growth rarely exceeds 15% (see Table 2 for historical rates of inflation and premium growth), we should expect the nominal wage constraint to bind rarely if ever for high-wage workers, but much more often for low-wage workers in periods of moderately low inflation (2–3%).[6]

2. Employee Contributions and the Tax Subsidy for Insurance

In any newspaper article on growing health insurance costs, the following remark by a former Xerox benefits manager is familiar rhetoric: "Employers will be increasing not only cost sharing in the premium paid each month but in every possible way, with either a co-payment or co-insurance."[2] The widely held view of non-economists is that employers bear the burden of premium hikes, but try to find ways to pass costs onto workers. Directly reducing wages, however, is not included under the heading "every possible way." The implication is that firms cannot recoup insurance costs through reduced wages, but can do so through other means—in particular, through the employee contribution, when firms require workers to pay directly some fraction of their premiums. Throughout the 1990's, firms required increasingly large employee contributions, which for most employers sacrifices preferential tax treatment (KFF, 2001). Employee contributions, though eligible for exemption in theory, contingent upon the satisfaction of certain federal regulations, are indeed taxed for nearly 75% of U.S. workers (Gruber, 2000). Given the tax subsidy to employer payments, the trend toward larger contributions is surprising and has been the subject of considerable theoretical discussion among economists.

Dranove, Spier and Baker (2000) present one common economic view: As more families have the option of multiple sources of coverage due to increasing female work force participation, each firm tries to avoid covering the whole family. Larger employee contributions make it less likely that workers will select family coverage. While plausible, this argument fails to explain why the contributions wouldn't be required only for family coverage, or why this trend would continue even as female labor force participation levels out. More generalized explanations along these lines contend that employee contributions are a means of filtering out any workers who do not value the benefit as much (Pauly, 1986). Following this logic, multiple factors have been shown to have at least a marginal role in the growth of employee contributions, include the growing prevalence of managed care, expanded eligibility of Medicaid for women and children, and cyclical conditions—which all may create an incentive for firms to direct workers towards cheaper options, on the margin (Gruber and McKnight, 2003). But fundamentally, this argument does not jibe with the standard incidence model – at least not without a key revision. If firms recoup the full cost of the benefit through decreased wages, then they should be indifferent between workers choosing individual plans, family coverage, or none at all: Wages will adjust to cover the average cost of all workers' insurance decisions. Thus, the Pauly/Dranove explanation implies some degree of wage rigidity that prevents this readjustment from happening, a point raised by Levy (1998) in her analysis of premium contributions. But if we are willing to assume some wage rigidity, then we first should consider whether this rigidity itself might lead to increasing premium contributions—even aside from the other factors discussed in the economics literature.

This section addresses this set of issues, asking: What if workers are unwilling to accept nominal wage cuts, but do not have similar attitudes towards their share of insurance premiums? In such cases firms will find it profitable to shift increasing insurance costs onto workers through premium contributions, even though this sacrifices a tax deduction, because the nominal wage constraint leaves them no other choice. This corresponds to the suggestion from industry that wage cuts simply are not in the firm's toolbox, but premium contributions are. Thus, this section explores the possibility that wage rigidity should be added to—and may in fact help justify—the list of explanations in the economics literature for the growth of premium contributions.

The revised model differs from the basic model in two ways. First, the firm's insurance expenditures in both periods are divided by $1+s$, where s is the subsidy to employer-provided insurance created by the federal tax exemption. As mentioned earlier, nearly 75% of U.S. workers pay taxes on their premium contributions—thus, throughout this section, I treat these contributions as taxable. Second, the Period 1 profit equation and total compensation constraint (WC1) include a term E for the employee contribution, which ranges between 0 and 1 as a share of the total premium. As we will see, the model's results hold even if E is positive in Period 0; setting $E_0 = 0$ for now is simply a normalization. Period 1 cost is the sum of the new wage and the inflated but subsidized insurance, minus the employee contribution.

$$\text{PERIOD 0} \quad \underset{Q_0}{\text{Max}} \quad \pi = p(Q_0) \cdot Q_0 - Q_0 \cdot [w_0 + b/(1+s)]$$

$$\text{PERIOD 1} \quad \underset{Q_1, w_1, E}{\text{Max}} \quad \pi = i \cdot p(Q_1) \cdot Q_1 - Q_1 \cdot [w_1 + b \cdot g \cdot i \cdot (1-E)/(1+s)]$$

$$\text{s.t.} \quad w_1/i + b \cdot g \cdot [1-E] \geq w_0 + b \quad \text{(WC1)}$$

$$\text{s.t.} \quad w_1 \geq w_0 \quad \text{(WC2)}$$

Note that the total compensation constraint (WC1) only includes the employer's portion of insurance as compensation, and it does so at the unsubsidized cost. This is because, by assumption, workers values insurance at the price it would cost them, rather than the firm's subsidized price. Thus, the subsidy s represents the tax wedge between the cost to the firm and the employee's perceived benefit. If anything, this is a conservative estimate of the employee's value of the benefit, since this is non-mandated insurance. If the worker did not desire insurance as part of compensation, both the worker and firm would be better off without it. Of course, if this tax wedge persisted at all levels of b, then firms would offer a large insurance benefit and zero wage. We can avoid this problem simply by fixing b, implying an upper limit to the generosity of insurance, motivated by its decreasing marginal utility.

When only WC1 binds, we obtain a familiar prediction—the firm should not require an employee contribution (see Appendix A for the derivation). The problem quickly degenerates into the situation outlined in Section 1, where no contribution was possible and w_1 absorbed the full added cost of insurance. But output may not be the same. The new first-order condition (FOC) yields:

$$p(Q_1) + p'(Q_1) \cdot Q_1 - [w_0 + b - gb + gb/(1+s)] = 0 \quad (3)$$

When $g = 1$, corresponding to no change in the real cost of insurance, Eq. (3) reduces to the identical FOC from Period 0, implying that $Q_0 = Q_1$. If none of the relative prices change, the firm should not change its optimal behavior. However, if $g > 1$, the firm's output will not be the same in Period 1. This result differs from the basic model, where increasing insurance costs had no effect on output, as long as the nominal wage constraint did not bind. The source of this difference is the distortion created by the tax subsidy. The firm's Period 1 marginal cost decreases relative to Period 0 if real compensation in Period 1 is smaller:

$$w_0 + b - gb + gb/(1+s) < w_0 + b/(1+s) \rightarrow g > 1$$

Thus, the firm's labor costs go down whenever the real cost of insurance increases. This illustrates the dynamic effect of the tax subsidy: If the subsidized portion of the firm's labor costs increases, then the firm will be able to produce more output than before and increase its profits. When the nominal wage constraint does not bind, firms actually benefit from growing health care costs, due to the tax preferential treatment of insurance. Of course, this benefit to the firms will be short-lived, if it exists at all, because labor markets will simply push the competitive compensation package to a new higher equilibrium. In this case, it will be the workers who benefit from the tax subsidy to insurance.

Given the firm's ability to shift costs onto workers through the employee contribution E, it will never be profit maximizing to offer total compensation above the competitive level. This rules out the situation in which WC2 binds but WC1 does not. Therefore, when WC2 binds, the optimal E^* satisfies WC1 as an equality, yielding:

$$E^* = [-(w_0/b) \cdot (r_i/i) + r_g]/g \tag{4}$$

We can predict when requiring a contribution will be optimal by solving for $E^* > 0$:

$$r_g > (w_0/b) \cdot (r_i/i) \qquad \text{Ineq. (2)}$$

Ineq. (2) simply restates the result from the base case, captured in Ineq. (1). The employer only needs to require an employee contribution when the nominal wage constraint binds. The tax subsidy does not change this critical threshold, since the subsidy only affects the firm's cost function, not the nominal wage constraint or the employee's valuation of total compensation.

This result offers a straightforward explanation for the phenomenon of employee contributions. Faced with restrictions on wages, firms are willing to sacrifice the preferential tax treatment of insurance payments in order to cut overall labor costs—just as the firms themselves contend. Equation (4) indicates that the factors that lead to increasing contributions are the same that made it more likely in the basic model for firms to bear some of the increased cost of insurance. E^* increases with growth in insurance costs (g) and size of benefits (b), and decreases with the cash wage (w) and general inflation (i). This relationship between employee contributions and inflation offers another testable hypothesis, considered in Section 3.

2.1. Effects on Price and Profits, with an Employee Contribution

When the nominal wage constraint binds, the firm requires an employee contribution. Substituting $w_0 = w_1$ and the expression for E^* into the maximization problem, we obtain a surprising result—the firm's optimal quantity and profit are independent of g:

$$\max_{Q_1, w_1} \pi = ip(Q_1) \cdot Q_1 - iQ_1[w_0 - (w_0/b) \cdot (r_i/i) + [b + (w_0) \cdot (r_i/i)]/(1+s)]$$

In essence, once the nominal wage constraint binds, the firm has a simple response to premium growth—shift the excess cost (beyond what the firm can extract from the worker's "inflation raise") onto the worker's contribution at the full unsubsidized price. A simple case illustrates this effect: Assume no inflation and a \$10 increase in per-unit (unsubsidized) insurance costs. The nominal wage constraint binds. The firm simply requires the employee to contribute \$10 towards insurance. The worker's real compensation is unchanged, and the firms' costs are unchanged, so output and profit are unchanged. The rational firm, with the flexibility to require a premium contribution, no longer needs to cut quantity or raise prices. This result may seem counterintuitive, but it is important to distinguish between the independence of profit with respect to growth in insurance costs *conditional* on the nominal wage constraint binding, and the strong effect of the growth rate g on the likelihood that the wage constraint binds in the first place.

More realistically, however, employee contributions are probably somewhat rigid in the short-term. The recent strike by California grocery workers illustrates this point, since the supermarkets' plan to require premium contributions for the first time was one of the workers' key complaints (NY Times, 2003). Any kind of stickiness in this element of compensation would push us back towards the basic model, where firms are unable to recoup the full increase in insurance costs. Thus, we might observe all of the effects discussed above—increased employee contributions, decreased firm profits, higher product prices, and reduced employment. When the nominal wage constraint binds, each element bears a share of the cost of rising health insurance premiums.

2.2. Generalizations

The model explored in this section makes two restrictive assumptions that bear more discussion. First, the firm requires no employee contribution at all in Period 0 and suddenly develops that option in the next period. This is simply a normalization. If we allow the Period 0 contribution share to be some $E_0 > 0$, then the additional Period 1 contribution share is $E_1(1 - E_0)$. The firm's decision is the same—it can stick with the current level of employee contribution (in which case $E_1 = 0$), or it can increase the contribution at the expense of the tax subsidy. The firm can simply ignore whatever portion of insurance the employee paid for in the previous year, because this will not affect either the competition constraint or the total cost of labor to the firm. This result resolves some of the model's apparent—and possibly troubling—implications regarding wages vs. contributions, such that new firms would never require an employee contribution, or that wage growth should always come first through reducing the contribution. In practice, new firms may start with some level of $E_0 > 0$, perhaps established for efficiency reasons of the kind discussed by Dranove, Spier and Baker (2000), and the same $E_0 > 0$ is likely to be the lower bound on a firm's required contributions even in times of wage growth. Neither of these possibilities changes the model's basic predictions.

This discussion points to a more general application of this model, which the basic model did not permit. In allowing the firm to set E, we are essentially allowing it to decrease the generosity of the benefit, b. The analysis in this section would be identical if instead of setting a contribution level E the firm was allowed to select any new $b_1 \leq gb_0$. Any compensation in the form of insurance is subsidized, so the firm will only decrease the level of benefits if it encounters the nominal wage constraint. Firms will sacrifice some of the tax subsidy if that is the only way to recoup higher premiums, and this can take the form of either increased employee contributions or less generous insurance—which in the extreme case could manifest itself as the firm dropping coverage completely. To the firm, these tools are identical. Thus, the dual patterns of increased employee contributions and decreased coverage generosity in recent years can both be explained by nominal wage stickiness.

The models in Sections 1 and 2 make two additional assumptions that need to be addressed: labor markets are perfectly competitive, and nominal wage cuts are infinitely costly to the firm. Appendix B presents a general model in which firms face a traditional upward-sloping labor supply and a finite cost of imposing a nominal wage cut (rather than a fixed constraint). The most important result is that the predictions of the models presented here persist under more general conditions—the results do not hinge on these two assumptions. The general model also predicts that firms facing greater costs from nominal wage cuts—perhaps due to high costs of worker turnover or difficulties in preventing shirking—will reduce wages less than other firms and bear more of the burden from growing premiums. Lastly, the general model predicts that wage stickiness will have a smaller impact on firms with more elastic labor supply, because its effect will be outweighed by more traditional labor supply concerns.

3. Empirical Analysis

In this section, I test the prediction that firms facing rapid premium growth are better able to shift those costs to wages if they are located in regions with high inflation and if their workers have high initial wages. Thus, low-income insured workers in regions of low inflation and high premium growth should: (A) have higher real wage growth; (B) experience greater increases in unemployment; and (C) face larger increases in premium contributions, when compared with similar workers in high-inflation regions.

In Section 1, we saw that the nominal wage constraint only binds in times of rapid premium growth and low general inflation. From 1991–2001, only the years 2000–2001 exhibited the three features needed for this identification strategy: low national inflation rates (3% or less), rapid premium growth (10% or more), and significant variation in regional inflation rates (BLS, 2002; Kaiser, 2001).[7] From March 2000 to March 2001, health insurance premiums grew at an average rate of 11%, while general inflation was 2.9%, and the regional inflation rate ranged from 2.3% in the South to 3.7% in the West (Kaiser, 2001; BLS, 2002).[8] While this absolute difference in inflation may seem small, it reflects a relative difference of 60%.

Furthermore, the effect we are looking for is nonlinear: According to Ineq. (1), firms facing 2.3% inflation in this period would be unable to shift rising premiums fully onto wages for workers whose health insurance was at least 28.5% of cash wages, while firms facing 3.7% inflation would face this problem for workers whose insurance represented more than 45% of their wages. Clearly, Southern firms would have been much more likely to encounter the constraint imposed by workers' aversion to nominal wage decreases.

Of course, the analysis would have more variation if it used state-level inflation, but the Bureau of Labor Statistics does not calculate state-level inflation rates, preferring the increased sample size and reliability of regional inflation estimates. This section tests for a wage response, an employment response, and an effect on employee premium contributions from 2000 to 2001. The model predicts that the interaction of employer-provided health insurance, low income, and low inflation should have positive effects on real wage growth, on the probability of losing one's job, and on the probability of an increased premium contribution. It is important to note that the model predicts that firms will *first* implement the wage cut as much as inflation permits, before adopting premium contributions, decreased benefits, or quantity-cutting, which disrupt the firm's prior profit-maximizing behavior. Thus, even if firms under low inflation are more likely to implement other changes, these should be in addition to a wage effect, rather than instead of it.

The data are from the March Income Supplement to the Current Population Survey (CPS), administered by the Department of Labor. The CPS uses a rotating nationally representative sample, in which a given household is included in the survey for 4 months, then off for 8 months, and then surveyed again for 4 more months. This means that at any point in time, half of the survey sample was surveyed 12 months earlier. The March Supplement contains detailed questions on hours worked, earnings, source of health insurance coverage, and demographic questions. The study sample contains all respondents who described themselves as working members of the labor force in both periods. In testing for employment effects, the sample is expanded to include those who were working in 2000 but out of work in 2001.

The model's predicted wage response is based on the hourly wage, rather than annual earnings. However, only a fraction of the CPS sample reports an hourly wage. Thus, imputation was required for the remainder of the sample, based on the annual earnings and total hours worked. Comparison of imputed values with observed wages for those who reported an hourly wage indicates that the imputation does not introduce any systematic bias (6.6% observed wage growth vs. 6.5% imputed). The final sample size was roughly 10,000 for analyses involving only the West and South, and 20,000 for regressions using the whole country.[9] Table 3 provides some key descriptive statistics for the four-region wage-regression sample.

3.1. Analysis & Results

3.1.1. Wage Effects.

3.1.1.1. Differences-in-Differences-in-Differences (DDD) with Sample Means. Comparing workers in March 2000 and March 2001, we can compare the mean difference in real

Table 3. Selected characteristics of the study sample.

Variable	Mean/percentage
Age	42.6
Sex	
Male	54.3%
Female	45.7%
Race	
White	87.8%
Black	7.8%
Asian/Pacific Islander	3.4%
Native American	1.0%
Schooling	
No high school diploma	9.8%
High school diploma	30.9%
Some college	29.9%
College graduate	29.4%
Region	
South	29.3%
Midwest	26.1%
West	24.0%
Northeast	20.6%
Employer-Provided HI (Year 0)	58.8%
Employer pays all	17.4%
Employer pays some	41.3%
Annual earnings	$36,614
# Observations (n)	*19,385*

wages between Southern and Western workers. The treatment group consists of workers with health insurance provided by their employers (as of March 2000[10]), who have a high school diploma or less. I use education to divide the sample because the model predicts an effect of inflation only for workers whose wages are small relative to their insurance benefit; but since wages are the dependent variable, sorting on that basis would be subject to bias. Year 0 wages and wage growth from Year 0 to Year 1 will be endogenously determined, and subject to regression to the mean. Thus, education serves as an exogenous proxy for wages, as has been done in similar labor analyses in lieu of stratifying by income (e.g. Eissa and Liebman, 1996).

There are two natural control groups for this treatment group of less-educated workers with health insurance—more highly-educated workers with employer-provided health insurance, and less-educated workers without employer-provided health insurance. Subtracting the mean change in wage among highly-educated workers with insurance filters out the standard compensating differential, thus addressing concerns about regional variation in health insurance growth. Subtracting the mean change in wage among less-educated workers without employer-provided insurance filters out any secular trend in lower-wage workers' incomes, as well as any differential growth rate across the two regions (including direct effects of inflation).

Table 4 summarizes the DDD analysis with sample means. In each group—the treatment group and the two controls—a pattern emerges that coincides with the model. In the treatment group, wage growth is higher in the South, just as the model predicts. There is little difference in the growth patterns between the two control groups across the two

Table 4. DDD Estimates of the wage effect of low inflation & high premium growth.

	Location	Year 2000	Year 2001	Time diff	% Change
		Treatment group—Low-education with health insurance			
($n = 1315$)	South	\$13.23	\$14.02	\$0.79 (\$.21)	5.97%
($n = 845$)	West	\$14.27	\$14.58	\$0.31 (\$.23)	2.17%
				$DD =$ \$0.48 (\$.31)	$D\% = 3.80\%$
		Insurance control group—High-education, with health insurance			
($n = 1954$)	South	\$20.37	\$22.34	\$1.97 (\$.30)	9.67%
($n = 1835$)	West	\$20.84	\$22.87	\$2.03 (\$.34)	9.74%
				$DD = -$\$0.06 (\$.45)	$D\% = -0.07\%$
		DDD for insurance control group			
		\$0.48		3.80%	
		− −\$0.06		− −0.07%	
		\$0.54 (\$.55)		3.87%	

	Location	Year 2000	Year 2001	Time diff	% Change
		Low-education control group—Low-education, no health insurance			
($n = 1266$)	South	\$9.68	\$10.43	\$0.75 (\$.17)	7.76%
($n = 906$)	West	\$9.78	\$10.54	\$0.77 (\$.31)	7.87%
				$DD = -$\$0.02 (\$.35)	$D\% = -0.11\%$
		DDD for low-education control group			
		\$0.48		3.80%	
		− −\$0.02		− −0.11%	
		\$0.50 (\$.47)		3.91%	

Note: Standard errors of the mean in parentheses.

regions. The fact that the higher-education group in both regions has higher wage growth than either low-education group likely reflects the trend of growing income inequality in recent years. That the treatment group's wage growth in both regions is lower than that of the low-education uninsured group indicates that there is still a *partial* wage offset in the treatment group due to premium growth, as we would expect. The total DDD estimates for the treatment group are \$0.54 (+3.87%) using the insured control group, and \$0.50 (+3.91%) using the low-education control group. The standard errors of these means are large, roughly of the same magnitude as the differences for the control groups; given the reliance on imputation and the small subsamples of each cell, the size of these errors is not surprising. To test for the significance of these effects controlling for demographics and using a larger sample as the baseline control group, we turn to the regression framework.[11]

3.1.1.2. Regression for Wage Effect. The dependent variable is the percentage change in person i's wage; X is a vector of demographic factors—age, race, and sex; and the I's are binary indicators for region (with *Northeast* omitted), high school diploma or less (*LowEd* = 1), and employer-provided health insurance in 2000 (*HI* = 1). The variable *Inf* is the regional inflation rate, equal to 2.3% for the South, 2.7% for the Midwest, 2.8% for the Northeast, and 3.7% for the West. The regression includes the full set of interaction terms for inflation, health insurance, and low-education. The inflation variable does not appear on its own, because the region dummies capture both the direct effect of regional inflation and any regional variation in wage trends. Thus, the remaining coefficients describe effects on *real* wages. The coefficient of interest, β_9, captures the interactive effect of low-education, health insurance, and regional inflation.

$$\begin{aligned}\%\Delta wage_i = {} & \alpha_0 + \beta_1 \cdot I_i^{\text{South}} + \beta_2 \cdot I_i^{\text{West}} + \beta_3 \cdot I_i^{\text{Midwest}} + \beta_4 \cdot I_i^{\text{HI}} + \beta_5 \cdot I_i^{\text{LowEd}} \\ & + \beta_6 \cdot Inf_i \cdot I_i^{\text{HI}} + \beta_7 \cdot Inf_i \cdot I_i^{\text{LowEd}} + \beta_8 \cdot I_i^{\text{HI}} \cdot I_i^{\text{LowEd}} \\ & + \beta_9 \cdot Inf_i \cdot I_i^{\text{LowEd}} \cdot I_i^{\text{HI}} + X_i \cdot \gamma + \varepsilon_i\end{aligned}$$

The model predicts that $\beta_9 < 0$: Higher inflation makes it more likely that these workers' wages will be fully docked for the increased premiums. The results are summarized in the first column of Table 5. All standard errors are clustered by region to adjust for the non-independence of observations within each region, necessary since in essence the regression has only four observations—the four regions (see Bertrand, Duflo and Mullainathan, 2002). The coefficient of interest is significantly negative, as predicted. At the margin, a 1% increase in regional inflation results in a *relative* decrease of 6.58% in the real wages of less-educated insured workers.[12] How does this coefficient compare with the model? The 6.58% estimate is larger than the direct effect of wage rigidity would predict on its own. For workers whose firms face the nominal wage constraint, a 1% increase in inflation relaxes the constraint by 1%, which the firm applies towards the cost of insurance. Thus, from this mechanism alone, we would expect a coefficient of −1.0, below the lower limit of the 95% Confidence Interval (−1.87 to −11.2). But this effect may be compounded if there is an accompanying quantity response of the kind described by Cutler and Madrian (1998): If firms facing the wage constraint use layoffs to concentrate hours among fewer employees, they have to pay higher wages to do so (assuming upward-sloping labor supply)—thus, *low* inflation may lead to

Table 5. Regression results of inflation effect on wages & unemployment.

Variable	% Change in wage	Pr(*Unemployed* | *employed* last year)
HI*LowEd*Inflation	**−6.58***	**−22.20***
	(1.48)	**(4.22)**
Age	−0.0018*	−.0055*
	(0.0005)	(0.0014)
Male	0.034	−0.175*
	(.012)	(.030)
White	−0.018	−0.120*
	(.013)	(.046)
South	0.015*	0.019
	(.002)	(.011)
West	−.032*	−0.116*
	(.004)	(−.018)
Midwest	−.037*	−0.013*
	(.001)	(.002)
Insurance (HI)	−0.197*	−.475*
	(.061)	(.171)
Low-education	−0.142*	−0.070*
(LowEd)	(.033)	(.007)
Inflation ∗ HI	3.90	9.02
	(1.92)	(5.37)
LowEd ∗ HI	0.20*	0.574*
	(.05)	(.162)
LowEd ∗ Inflation	3.27	10.87*
	(1.05)	(0.13)
# Observations (n)	*19,385*	*21,271*

Notes: Standard errors are given in parentheses, clustered for the non-independence of the variables within the four regions.
*Indicates significance at $p \leq .05$.

wages even higher than expected, making the coefficient on the inflation / treatment-group interaction term even more *negative* than expected. Taking the direct wage effect along with this quantity effect, the coefficient is within a reasonable range of the model's prediction. Next, we test for this hypothesized employment change.

3.1.2. Employment Response.

3.1.2.1. Differences-in-Differences (DD) with Sample Mean Hazard Rates. The framework described above for testing the wage response can be applied similarly to test for employment changes. The only difference is that the analysis now requires a binary dependent variable—the hazard rate of unemployment, Pr*(Unemployment | Employment the previous year)*. Employment is based on the respondent's self-reported labor status in 2001.

Respondents who describe themselves as active in the labor force are considered employed, and all other respondents are considered unemployed. Note that the entire sample was employed in 2000, so this method does not falsely label children or the permanently disabled as unemployed. Table 6 summarizes the DD estimates of unemployment hazard rates in the South vs. the West, in the treatment group vs. the two controls. The only significant location difference is in the treatment group. For this group, living in the South increases the unemployment hazard rate by 2.18%, consistent with the model. The negligible location differences for the two control groups support the identifying assumption that variation in inflation across regions should not create differential labor market trends in groups unaffected by the nominal wage constraint.

Table 6. DD Estimates of the effect of low inflation & high premium growth on unemployment hazard rates.

	Location	Hazard Rate
Treatment group—Low-education, with health insurance		
(n = 1444)	South	8.80% (0.75%)
(n = 906)	West	6.62% (0.83%)
Location difference = 2.18% (1.12%)		
Insurance control group—High-education, with health insurance		
(n = 2110)	South	6.26% (0.53%)
(n = 1968)	West	5.79% (0.53%)
Location difference = 0.47% (0.75%)		
DD for insurance control group		
		2.18%
		−0.47%
		1.71% (1.35%)

	Location	Hazard Rate
Low-Eucation control group—Low-education, no health insurance		
(n = 1428)	South	13.03% (0.88%)
(n = 1049)	West	13.30% (1.06%)
Location difference = −0.18% (1.38%)		
DD for low-education control group		
		2.18%
		− −0.18%
		2.36% (1.78%)

Note: Standard errors of mean in parentheses.

3.1.2.2. Regression for Job Loss. This regression uses the same specification as the wage regression, except for the change to a binary dependent variable and the probit form. The second column in Table 5 reports these results. The coefficient of interest is on *Health Insurance, Low-Education, & Inflation*, and it is strongly negative, as predicted. Taking the coefficient's slope at its mean value, an added point of inflation produces a 3.33% decrease in the risk that a worker in the treatment group becomes unemployed. Unlike the wage effect, there is no way to test the magnitude of this estimate against the model, since the degree of quantity change will depend on the elasticity of output demand and the firm's ability to concentrate labor among fewer workers. The most we can say is that an estimate of 3.33% does not seem patently unreasonable.

3.1.3. Employee Premium Contributions.

3.1.3.1. DD with Sample Mean Hazard Rates. The analysis of premium contributions is limited by the fact that the only relevant data in the CPS come from the item: "Does your employer pay for all, some, or none of your health insurance?" The group of interest consists of workers who switched from "All" in 2000 to "Some" in 2001. The model predicts that this change is more likely for workers in lower-inflation regions. Defining the variable *Contribution* as 1 if a worker moves from "All" to "Some," and 0 otherwise, we can calculate the hazard rate by taking the mean of *Contribution* among workers who reported "All" in 2000 and were still employed and insured in 2001 ($n = 2820$, roughly 15% of the original sample). Taking this hazard rate separately for highly-educated and less-educated workers, in the South vs. the West, provides a DD estimate of the effect of inflation on the likelihood of increased employee contributions. Table 7 shows that the hazard rate is higher for Southern

Table 7. DD Estimates of the hazard rate for new premium contributions.

	Location	Hazard Rate
Treatment group—Less educated workers		
($n = 228$)	South	50.4% (3.3%)
($n = 220$)	West	45.0% (3.4%)
Location difference = 5.4% (4.7%)		
Control group—Highly educated workers		
(n=431)	South	44.5% (2.4%)
(n=661)	West	43.5% (2.1%)
Location difference = 1.0% (3.2%)		
Difference-in-difference		
5.4%		
−1.0%		
4.4% (5.7%)		

Note: Standard errors of mean in parentheses.

firms with less-educated workers. Again, the raw means are consistent with the model but have large standard errors, so we turn to the regression analysis.

3.1.3.2. Regression for Requiring Contributions. I specified the following probit regression:

$$\Pr(Contribution_i = 1) = \alpha_0 + \beta_1 \cdot I_i^{\text{South}} + \beta_2 \cdot I_i^{\text{West}} + \beta_3 \cdot I_i^{\text{Midwest}} + \beta_4 \cdot I_i^{\text{LowEd}} + \beta_5 \cdot Inf_i \cdot I_i^{\text{LowEd}} + X_i \cdot \gamma + \varepsilon_i$$

The coefficient of interest is β_5, which captures the interactive effect of low-education and inflation. *HI* is no longer a right-hand side variable, since insurance status is now the dependent variable. The model predicts that $\beta_5 < 0$: Firms facing higher general inflation are better able to shift insurance costs onto wages, so they are less likely to start requiring an employee contribution. Table 8 summarizes the results. The coefficient of interest is negative as predicted and statistically significant, with a slope estimate of −2.82% per point of inflation. Again, the model does not provide a ready test of this estimate's plausibility. But it is worth noting that the employment and premium contribution effects

Table 8. Probit regression results for contribution hazard rates.

Variable	Pr (Contribution in 2001 | full coverage in 2000)
LowEd ∗ Inflation	**−7.19*** **(2.60)**
Age	−0.0052* (0.0017)
Male	−0.026 (.021)
White	−0.239* (.044)
South	0.070* (.009)
West	0.034* (.005)
Midwest	−0.104* (.004)
Low-education (LowEd)	0.295* (.085)
# Observations (n)	*2820*

Notes: Standard errors are given in parentheses, clustered for the non-independence of the variables within the four regions.
*Indicates significance at $p \leq .05$.

indicate that each bears a portion of the incidence of rising insurance costs. Firms are unable to recoup their losses fully through premium contributions and must turn to quantity cutting, suggesting that contributions—though less rigid than wages—are not fully flexible either.

3.2. Variation in Inflation: A Validity Check

An obvious objection to the identification strategy is that the inflation rate is far from exogenous and may be correlated with economic outcomes in a way that casts doubt on the results described above. The DDD framework addresses many of these concerns. But an additional way to verify that these effects are in fact due to nominal wage stickiness and rising insurance premiums is to conduct the identical analysis during a two-year period in which regional inflation rates varied but insurance premiums were *not* growing rapidly.

Table 9. Comparing Rapid Premium Growth (2000–2001) vs. Slow Premium Growth (1998–1999).

Variable	1998–1999	2000–2001
Single year regressions		
	% Change in wage	
HI ∗ LowEd ∗ Inflation	−3.50	−6.58*
	(3.74)	(1.48)
	Pr(Job Loss)	
HI ∗ LowEd ∗ Inflation	11.75	−22.20*
	(16.77)	(4.22)
	Pr(Contribution)	
LowEd ∗ Inflation	44.44*	−7.19*
	(6.05)	(2.60)
Combined two-year regressions		
	% Change in Wage	
HI ∗ LowEd ∗ Inflation ∗ yr00	−3.20	
	(3.76)	
	Pr(Job Loss)	
HI ∗ LowEd ∗ Inflation ∗ yr00	−33.97*	
	(15.97)	
	Pr(Contribution)	
LowEd ∗ Inflation ∗ yr00	−48.56*	
	(6.70)	

Notes: Standard errors are given in parentheses, clustered for the non-independence of the variables within the four regions for the single-year regression, and clustered within the eight region-years for the two-year regressions.
*Indicates significance at $p \leq .05$.

Under these circumstances, the model would predict no wage or employment effect, but any argument about direct effects of differential inflation should still apply. CPS data from 1998 and 1999 were chosen for this analysis due to their proximity to the study period and their regional variation in inflation. During this 12-month period, average insurance premiums grew by only 4.8%, while general inflation rates were similar in magnitude and variation as during the study period—ranging from 2.0% in the South to 2.7% in the West. The results are summarized in Table 9. The coefficients of interest for 1998–1999 are non-significant for wage and employment effects at the $p \leq .20$ level, but the premium contribution effect is significant in the opposite direction as in the 2000–2001 regression—and almost implausibly high. This raises concerns that the premium contribution estimate is not as robust as the wage and employment measures, due to the limitation of the "all, some, or none" measure.

Combining these two years into the same regression allows us to assess more directly the model's predictions, with the year ('98–'99 vs. '00–'01) now used as a fourth difference. The full set of interaction terms is included, as well as region-year fixed effects, and standard errors are clustered by region-year. These results are also in Table 9. The coefficient of interest for the wage regression is of similar magnitude as in the single-year '00–'01 regression, but the standard error is larger. Thus, the wage coefficient is not statistically significantly less than 0, but its confidence interval still includes the estimate from the '00–'01 regression. The coefficients for the employment and employee contribution are both significantly negative, as the model predicts. These two-year results offer moderate support for the conclusion that wage stickiness is the basis for the '00–'01 results. More specifically, the data support the model's overall predictions, but cast some doubt on the validity of the "all, some, or none" premium contribution measure.

4. Conclusion

If workers resist nominal wage cuts in a way that is costly to their employers, the standard incidence model of employer-provided health insurance may not completely capture the effects of rapid health insurance premium growth. A simple model of nominal wage stickiness goes a long way towards resolving two apparent paradoxes—employers' claims that growing health care costs are a burden to them, and the increasingly widespread use of employee premium contributions despite their tax disadvantages.

Empirical analysis offers support for several of the model's predictions: Less-educated and insured workers in low inflation regions should have experienced higher real wage changes, an increased likelihood of unemployment, and an increased likelihood of a required premium contribution, relative to their counterparts in high-inflation regions. I found consistent evidence for the wage and employment effects, and less robust but suggestive evidence regarding premium contributions. Overall, the regression results support the model's predictions, although an alternative data set with information on the size of employee premium contributions would allow for more convincing analysis of that particular effect.

What besides the proposed model could account for these findings? The use of two distinct control groups—highly-educated insured workers, and less-educated uninsured workers—rules out many of the plausible alternatives. Any alternate explanation must explain why wage growth, employment, and premium contributions for low-income insured workers were all affected differentially than for insured workers in general, or low-income workers in general. For instance, one might argue that the West experienced greater premium growth during this period, so the resulting wage offset was greater in the West than in the South. But the general effect of employer-provided insurance is filtered out by the DD estimate for the highly-educated insured workers. Similar logic regarding the other control group rules out the possibility that the observed effect is due to the Southern economy differentially favoring low-wage workers. Lastly, concerns that firms' offering of insurance itself might respond to differential inflation would need to explain why this process only affects less-educated workers, and why it did not produce an analogous effect in the comparison period of 1998–1999.

Of course, there are several limitations of this empirical analysis that should be mentioned. As discussed earlier, the CPS item on employee premium contributions is far too imprecise to test conclusively the model's predictions on this parameter. A dataset with a continuous range of premium contributions (either on a percentage or absolute dollar basis) would allow for much more sensitive analysis of this issue.

More generally, the identification approach used in this paper for all three measured effects—wages, employment, and premium contributions—suffers from two primary limitations: First, by selecting variation in regional inflation rates as the identification strategy, we are significantly limited in terms of years for analysis. As mentioned earlier, the combination of low general inflation, high premium growth, and significant regional variation in inflation is relatively uncommon over the last 15 years. Alternative approaches could therefore yield larger sample sizes. Secondly, when using a household survey such as the CPS, the empirical analysis must focus on workers rather than firms, though the model's predictions are primarily at the level of the firm. Firms make decisions based on their average worker, which means that an analysis of individual workers' wages and benefits will be biased if there is significant wage and benefit heterogeneity in the workforce of a given firm.

What future research, then, could address some of these limitations and test additional implications of the model? Firm-level data would be an important first step in addressing the potential impact of worker heterogeneity. For instance, such data could ascertain: Do low-wage workers at firms that also have high-income workers fare differently in times of rapid premium growth than workers at exclusively low-income firms? Furthermore, firm-level data with detailed longitudinal tracking of wages and benefits would allow us to test more directly for nominal wage rigidity: Are wages more likely to be stagnant in years of rapid premium growth, and are premium contributions more likely to increase in years when wages are stagnant as opposed to increasing (either due to inflation or productivity gains)? Another important implication of the model not tested here is the effect of firm size: Facing similar premium growth and employing a similar workforce, the model predicts that smaller firms would be more likely to reduce benefit generosity or require larger premium contributions. Overall, the results in this paper are highly suggestive of a role for wage rigidity in explaining firms' response to health insurance premium growth, and future research—ideally with firm-level data—could more precisely quantify the magnitude of this effect and examine some of the model's as-yet untested implications.

In conclusion, empirical analysis offers support for key predictions of this paper's model, indicating that wage stickiness partially redistributes the incidence of rapidly growing insurance premiums. Rather than coming solely out of workers' wages, as the traditional incidence model predicts, escalating health care costs force firms, newly unemployed workers, and premium-paying employees to bear some of the burden of employer-provided health insurance.

Appendix A: Optimization & Comparative Statics for Sections 1 and 2

A.1. The Basic Model

Assuming concave profit, the FOC identifies a unique optimum.

$$\text{FOC: } p(Q_0) + p'(Q_0) \cdot Q_0 = w_0 + b \tag{A1}$$

In Period 1, the FOC for profit maximization yields:

$$i \cdot p(Q_1) + i \cdot p'(Q_1) \cdot Q_1 = (w_1 + g \cdot i \cdot b)$$

When only WC1 binds, we are left with the same FOC as Period 0, so $Q_1 = Q_0$.
When only WC2 binds, $w_1 = w_0$. The FOC yields:

$$p(Q_1) + p'(Q_1) \cdot Q_1 = (w_0/i + g \cdot b) \tag{A2}$$

This differs from the Period 0 FOC (Eq. (A1)) only on the cost side. By construction, we know that $w_0/i + g \cdot b > w_0 + b$, so the firm's real marginal cost has increased from Period 0.

When both constraints bind, $w_1 = i(w_0 + b - gb)$, and $w_1 = w_0$. Substituting, we get $g = 1 + w_0(i - 1)/bi$. Let $g = 1 + r_g$ and $i = 1 + r_i$. This yields: $r_g = (w/b) \cdot (r_i/i)$. From this equation, it is simple to verify that the nominal wage constraint binds when:

$$r_g \geq (w/b) \cdot (r_i/i) \qquad \text{Ineq. (1)}$$

A.1.1. Comparative Statics. Let Eq. (A2) serve as an implicit function relating p, Q, b, g, and i.

$$\frac{dp}{dg} = \frac{p'(Q) \cdot b}{p'(Q) + p''(Q) \cdot Q + p'(Q)} = \frac{b}{2 + E_{QQ}} \tag{A3}$$

$$\frac{dp}{di} = \frac{-p'(Q) \cdot i^{-2} w_0}{p'(Q) + p''(Q) \cdot Q + p'(Q)} = \frac{-w_0}{i^2(2 + E_{QQ})} \tag{A4}$$

$$\frac{d\pi}{dg} \bigg/ i = \frac{1}{i}\left(\frac{\partial \pi}{\partial Q} \cdot \frac{dQ}{dg} + \frac{\partial \pi}{\partial g}\right) = -Q_1 b \tag{A5}$$

E_{QQ} is the elasticity of the slope of demand. Equation (A3) shows that firms respond to growing insurance costs by increasing prices, as long as $E_{QQ} > -2$; this is not an overly restrictive assumption, given that $E_{QQ} < -2$ represents an extraordinarily rapidly changing elasticity of demand and is likely to be quite uncommon. The size of the price change is proportional to b, and is even larger if E_{QQ} is negative (i.e. if demand elasticity is increasing in Q). Equation (A4) indicates that inflation has a negative effect on the real price change, again as long as $E_{QQ} > -2$. High inflation is a buffer against the nominal wage constraint, as a function of w_0.

Unlike the price function, profit is responsive to inflation, so the derivative in Eq. (A5) is divided by i to estimate changes in real profit in Period 1. The firm loses money, in direct proportion to the quantity of labor employed and the size of the health insurance benefit.

$$\frac{d\pi^*}{di} = \frac{Q_1 w_0}{i^2} \tag{A6}$$

The derivative for real profit (π^*), Eq. (A6), is non-negative. When WC2 holds, the firm's real profits increase with inflation, since inflation provides a buffer against premium increases.

A.2. Employee Contributions

When WC1 binds, the maximization problem is:

$$\max_{Q_1, E} \pi = ip(Q_1) \cdot Q_1 - iQ_1[w_0 + b - gb + gbE + gb(1-E)/(1+s)]$$

$$\frac{\partial \pi}{\partial E} = -iQ_1 \cdot gb\left(1 - \frac{1}{1+s}\right)$$

The term in parenthesis is positive, so the whole derivative is negative. Profit is always decreasing in E, under WC1. Thus, we will get a corner solution at $E = 0$.

When will $E \geq 1$? When $-1 \geq (w_0/b) \cdot (r_i/i)$. Thus, $E = 1$ only under extreme deflation ($i \ll 1$, *implying* $r_i \ll 0$) and extremely high wages relative to insurance. These extreme circumstances allow us to ignore the constraint that $E < 1$ and focus on the interior solution.

Appendix B: The General Model

The firm faces an unconstrained optimization. $L(w+b)$ is labor supply, with $L'(w+b) > 0$. L is normalized to the amount of labor needed per unit output, so $L = Q$. For simplicity, normalize Period 0 benefit $b = 0$. In Period 0, the firm optimizes:

$$\max_{w_0} \pi = p(L(w_0)) \cdot L(w_0) - L(w_0) \cdot w_0$$

$$\text{FOC}: w_0 = \frac{P \cdot E_{\text{LS}} \cdot (E_D + 1)}{E_{\text{LS}} + 1} \tag{B1}$$

where $E_{\text{LS}} = (\partial L/\partial w) \cdot (w/L)$ & $E_D = (\partial p/\partial Q) \cdot (Q/p)$.

In Period 1, an exogenously determined health benefit h is added (given the Period 0 normalization, h is the added nominal cost of insurance). Let $C(-)$ be the firm's cost function for a nominal wage decrease. Defining $\Delta w = w_1 - w_0$, then $C(\Delta w) \geq 0; C(\Delta w) = 0$ if $\Delta w \geq 0$. For $\Delta w < 0$, then $C'(\Delta w) < 0$, and $C''(\Delta w) > 0$. Thus, both total cost and marginal cost get larger as Δw becomes more negative. In Period 1, labor supply responds to the real value of total compensation. Assuming that profit is strictly concave, there is a unique optimum:

$$\operatorname*{Max}_{w_1} \pi = i \cdot p\left[L\left(\frac{w_1 + h}{i}\right)\right] \cdot L\left(\frac{w_1 + h}{i}\right) - L\left(\frac{w_1 + h}{i}\right) \cdot (w_1 + h) - C(w_1 - w_0) \tag{B2}$$

$$\text{FOC}: \frac{w_1 + h}{i} = \frac{E_{\text{LS}} \cdot \left[P \cdot (E_D + 1) - \frac{C'}{L'}\right]}{E_{\text{LS}} + 1} \tag{B3}$$

Comparing Eqs. (B1) and (B3), it is clear that total real compensation stays the same if $C'/L' = 0$. One way for this to occur is if L' goes to infinity—perfectly elastic labor supply. Or, C' could be zero, which occurs only if $(w_1 - w_0) \geq 0$. If the firm can keep total real compensation the same without a nominal wage decrease, it will do so. This is possible as long as:

$$\frac{w_1 + h}{i} = w_0 \quad \text{and} \quad w_1 \geq w_0.$$

These are identical to the wage constraints from the basic model. Thus, Ineq. (1), which indicates when the nominal wage constraint will bind (or, when $C(\Delta w) > 0$), holds generally.

When a firm must implement a nominal wage cut, the following FOC serves as an implicit function for comparative statics (for notational ease, the arguments of the functions are omitted):

$$0 = p' \cdot L' \cdot L + L' \cdot p - L - L' \cdot (w_1 + h)/i - C' \tag{B4}$$

We can equivalently examine the wage, output, or price—these all define one another.

B.1. Effect of Health Insurance Costs on Price

$$\frac{dp}{dh} = \frac{\partial p}{\partial L} \cdot \frac{\partial L}{\partial w} \cdot \frac{dw}{dh} + \frac{\partial p}{\partial h} = (p'L'/i) \cdot \left[-\frac{\partial FOC}{\partial h} \bigg/ \frac{\partial FOC}{\partial w}\right] + p' \cdot L'/i,$$

where

$$\frac{\partial FOC}{\partial h} = p'' \cdot (L')^2 L/i + p'L \cdot L''/i + 2p' \cdot (L')^2/i + pL''/i - 2L'/i - L'' \cdot (w_1 + h)/i^2$$

$$\frac{\partial FOC}{\partial w} = p'' \cdot (L')^2 L/i + p'L \cdot L''/i + 2p' \cdot (L')^2/i + pL''/i - 2L'/i - L'' \cdot (w_1 + h)/i^2 - C''$$

When the firm does not implement a nominal wage cut, $C'' = C = 0$, and the two partial derivatives of the FOC are equal. Thus, $dp/dh = 0$, and premium growth does not affect the firm. But if Ineq. (1) holds, $C'' > 0$, and since both partial derivatives of the FOC are negative by the concavity of profit, $[\partial FOC/\partial h]/[\partial FOC/\partial w] < 1$. Overall, $dp/dh > 0$. Facing the cost of the nominal wage cut, the firm responds to premium growth by raising its prices, as in the base case.

B.2. Effect of Inflation on Price

$$\frac{dp}{di} = \frac{\partial p}{\partial L} \cdot \frac{\partial L}{\partial w} \cdot \frac{dw}{di} + \frac{\partial p}{\partial i} = (p'L'/i) \cdot \left[-\frac{\partial FOC}{\partial i} \Big/ \frac{\partial FOC}{\partial w} \right] + p' \cdot L' \cdot \left(-\frac{w_1 + h}{i^2} \right), \text{ where}$$

$$\frac{\partial FOC}{\partial i} = \left(-\frac{w_1 + h}{i^2} \right) \cdot \left[p'' \cdot (L')^2 L + p'L \cdot L'' + 2p' \cdot (L')^2 + pL'' - 2L' - L'' \cdot \frac{w_1 + h}{i} \right]$$

$$\frac{\partial FOC}{\partial w} = \frac{1}{i} \cdot \left[p'' \cdot (L')^2 L + p'L \cdot L'' + 2p' \cdot (L')^2 + pL'' - 2L' - L'' \cdot \frac{w_1 + h}{i} \right] - C''$$

When $C'' = 0$, the two lengthy partial derivatives are equivalent, except for multipliers of $-(w_1 + h)/i^2$ in $\partial FOC/\partial i$ and $1/i$ in $\partial FOC/\partial w$, but these cancel out overall and yield $dp/di = 0$. The optimal real price remains the same in Period 1 if Ineq. (1) doesn't apply. When the firm implements a nominal wage cut, $\partial FOC/\partial w$ becomes more negative, so *dp/di* < 0. Thus, as in the basic model, the degree of price distortion compared to the Period 0 optimum decreases with inflation. Higher inflation also means less distortion in the wage and less of a decrease in profit.

B.3. Effect of the Nominal Wage Cost Function on Wages

Parameterize the cost function as

$$C(x) = \left\langle \frac{k \cdot x^2 \quad \text{if} \quad x < 0}{0 \quad \text{if} \quad x \geq 0} \right. .$$

Now k can be used for comparative statics, where a larger k indicates a steeper cost of imposing a nominal wage cut:

$$\frac{dw}{dk} = \frac{2(w_1 - w_0)}{[p'' \cdot (L')^2 L/i + p'L \cdot L''/i + 2p' \cdot (L')^2/i + pL''/i - 2L'/i - L'' \cdot (w_1 + h)/i^2 - 2k}$$

$$\text{TERMS}: \quad A \quad + \quad B \quad + \quad C \quad + \quad D \quad - \quad E \quad - \quad F \quad - \quad G$$

This derivative is positive. When the nominal wage cost is imposed, $w_1 - w_0 < 0$, so the numerator is negative. The denominator is $\partial^2\pi/\partial w^2$ (from the implicit function theorem) and therefore negative by the concavity of profit. Firms with a larger k will have higher Period 1 wages in order to minimize the nominal wage cut. Delving into the denominator, terms C, $-E$, and $-G$ are all negative, while the signs of the second-order terms (A, B, D, F) are ambiguous. If P'' is positive, then the whole derivative becomes larger—a firm facing increasingly elastic demand will be more likely to keep its wages high to maintain output and not drive away price-sensitive customers. If L' is highly positive, terms C and E increase, and the whole derivative gets closer to zero: Firms facing more elastic labor supply are less concerned with the nominal wage decrease and more concerned with the labor supply response to total compensation.

Acknowledgments

This paper was written with the support of a fellowship from the National Science Foundation. Many thanks to Joe Newhouse, David Cutler, Caroline Hoxby, and Richard Freeman for helpful comments and advice, and to Melissa Wachterman for being an excellent sounding board. Of course, any errors or flaws in the paper are solely mine.

Notes

1. J. Thottam (2002). "Business, Health Thyself." *Time Magazine*, 14 Oct 2002, Bonus Section: Inside Business.
2. J. Kendall (2001). "The Health Insurance Malady." *BusinessWeek Online,* 5 March 2001.
3. The combination of downward-sloping demand and competitive labor supply may seem unusual, but it is an accurate description of many service industries with unskilled labor. Imagine multiple fast-food franchises in a given neighborhood: Compensation for labor is essentially competitive, but differentiated demand across restaurants makes monopolistic competition a reasonable approximation for product markets.
4. With this decrease in labor supply, the assumption of a competitive labor market becomes unreasonable; Appendix B replaces this assumption with a more generalized labor supply function.
5. Fuchs (1994) recognizes the relevance of low-wage vs. high-wage workers; in fact, he sees it as the *only* issue in whether a firm can "afford" coverage. But, in ignoring the synergistic effect of firm size and wages, this view fails to explain why low-wage workers at large firms are much more likely to have insurance than those at small firms.
6. Adding in the cost to the employer of the Medicare Part A payroll tax may increase the likelihood of the nominal wage constraint binding under times of rapid premium growth: The magnitude of this payroll tax has historically varied with the rate of health care cost growth, either through Congress directly increasing the rate or through increasing the taxable base of payroll per worker. Thanks to Joseph Newhouse for this insight.
7. The empirical measurement of premium growth is similar but not identical to the theoretical parameter g in the model, since the empirical rate of growth also includes any changes in the size or scope of the average

employer-provided insurance plan, whereas g in the model only reflected premium growth for a given level of benefits.

8. The regional inflation rates are the region-specific ratios of the 12-month average price indices for 2000 and 2001.
9. The samples from 2000 and 2001 were linked using household and personal identification numbers. Any matched file that indicated a change in age of more than two years or a change in sex was dropped ($<5\%$ of the sample). The sample also excluded workers reporting zero income for either year and those with implausible wage changes ($>400\%$, or $<-80\%$; i.e. five-fold changes in either direction), which closer investigation indicated were nonsensical typographical errors. In both time periods, nominal imputed wages were constrained by the \$5.15 minimum wage in place throughout the study period.
10. Previous empirical work comparing the CPS estimates on insurance coverage to other national surveys supports interpreting CPS items as point-in-time estimates, rather than coverage over the previous year (Swartz, 1986).
11. In a regression for South and West, the coefficient of interest significantly differs from zero in the expected direction, but these results are omitted here in favor of the more robust four-region specification.
12. This change is only relative. In absolute terms, the 6.58% decrease is balanced against the 3.90% increase of *Inflation & Insurance* and 3.27% increase of *Low-Education & Inflation* ($-6.58\% + 3.90\% + 3.27\% = +0.60\%$).

References

Bertrand, M., E. Duflo and S. Mullainathan. (2002). "How Much Should We Trust Differences-in-Differences Estimates?" *NBER Working Paper 8841*. Cambridge MA: NBER.

Bureau of Labor Statistics. (2002). *Consumer Price Indices*. Available at: <www.bls.gov/cpi/>, accessed March 20, 2002.

Card, D. and D. Hyslop. (1997). "Does Inflation 'Grease the Wheels' of the Labor Market?" In C.D. Romer, D.H. Romer (eds.) *Reducing Inflation: Motivation and Strategy.* NBER Studies in Business Cycles, vol 30. Chicago & London: Univ. Chicago Press, pp. 71–114.

Current Population Survey, March Supplement. (2000). Available at: <ferret.bls.census.gov/macro/032001/perinc/new01_016.htm>, accessed January 6, 2002.

Cutler, D.M. and B.C. Madrian. (1998). "Labor Market Responses to Rising Health Insurance Costs: Evidence on Hours Worked." *RAND Journal of Economics* 23(3), 509–530.

Dranove, D., K.E. Spier and L. Baker. (2000). "Competition' Among Employers Offering Health Insurance." *Journal of Health Economics* 19, 121–140.

Eissa, N. and J.B. Liebman. (1996). "Labor Supply Response to the Earned Income Tax Credit." *Quarterly Journal of Economics* 111(2), 605–637.

Fuchs, V. (1994). "The Clinton Plan: A Researcher Examines Reform." *Health Affairs* Spring(I), 102–114.

Gruber, J. and A. Krueger. (1991). "The Incidence of Mandated Employer-Provided Insurance: Lessons from Workers' Compensation Insurance." In *Tax Policy and the Economy.* Cambridge: MIT Press for NBER, pp. 111–143.

Gruber, J. R. McKnight. (2003). "Why Did Employee Health Insurance Contributions Rise?" *Journal of Health Economics* 22(6), 1085–1104.

Gruber, J. (1994). "The Incidence of Mandated Maternity Benefits." *American Economic Review* 84(3), 622–641.

Gruber, J. (2000). "Health Insurance and the Labor Market." In A.J. Cuyler and J.P. Newhouse (eds) *The Handbook of Health Economics*. Amsterdam: North Holland, pp. 645–706.

Kahn, S. (1997). "Evidence of Nominal Wage Stickiness from Microdata." *American Economic Review* 87(5), 993–1008.

Kahneman, D., J.L. Knetsch and R. Thaler. (1986). "Fairness as a Constraint on Profit Seeking: Entitlements in the Market." *American Economic Review* 76(4), 728–741.

Kaiser Family Foundation [KFF]. (2001). "Employer Health Benefits: 2001 Summary of Findings." Available at: <www.kff.org/content/2001/20010906a/EHB2001_sof.pdf>, accessed August 1, 2003.

Kaiser Family Foundation [KFF]. (2002). "Employer Health Benefits: 2002 Summary of Findings." Available at: <www.kff.org/insurance/3251-index.cfm>, accessed July 15, 2004.

Krueger, A.B. and U.E. Reinhardt. (1994). "The Economics of Employer Versus Individual Mandates." *Health Affairs* Spring (II), 34–53.

Levy, H.G. (1998). "Employee Contributions to Health Insurance Premiums, Part Two: Sticky Wages or Risk Sharing." Chapter 4. In *Employer-Sponsored Health Insurance: Costs, Coverage, and Contributions*. Unpublished Doctoral Dissertation, Princeton University.

New York Times. (2003). "Two Sides Seem Entrenched in Supermarket Dispute." By S Greenhouse. 10 November 2003, p. 12.

Pauly, M.V. (1986). "Taxation, Health Insurance, and Market Failure in the Medical Economy." *Journal of Economic Literature* 24(2), 629–675.

Pauly, M.V. (1997). *Health Benefits at Work*. Ann Arbor: University of Michigan.

Summers, L.H. (1989). "Some Simple Economics of Mandated Benefits." *American Economic Review* 79(2), 177–183.

Swartz, K. (1986). "Interpreting the Estimates from Four National Surveys of the Number of People Without Health Insurance." *Journal of Economic and Social Measurement* 14(3), 233–243.