Federal Reserve Bank
of Chicago

First Quarter 2008

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.

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

chicagofed.org

ISSN 0164-0682

Contents

First Quarter 2008, Volume XXXII, Issue 1

Obesity, disability, and the labor force

Kristin F. Butcher and Kyung H. Park

Introduction and summary

In this article, we investigate how the rise in obesity over the past three decades is related to non-employment. In recent years, unemployment rate figures—joblessness among those actively seeking work—have been low by historical standards. At the same time, however, there has been a rise in the fraction of men who are not actively seeking work.[1] The labor force participation of men of prime working age is low by historical standards, and this has coincided with an expansion in the Social Security Disability Insurance (SSDI) program.

A number of researchers studying the increase in men's non-employment have pointed out that it takes place against a backdrop of improving health (Juhn, Murphy, and Topel, 2002; and Autor and Duggan, 2003). However, these improvements in health are typically measured by mortality rates, which have been declining over time (Cutler and Richardson, 1997). Obesity rates, on the other hand, have climbed dramatically during the past 30 years. To put the increase in perspective, the median male in 2002 would have been heavier than 75 percent of the male population in 1976, using a body mass index (BMI) distribution.

There are a number of reasons that increases in obesity might be linked to decreases in employment. Increases in obesity might affect the ability to work—for example, obese people are more likely than others to have health problems—or the willingness to work, depending on the availability of alternatives to working. We call these "supply side" factors—those factors that affect whether or not an individual is willing and able to take a job. There may also be "demand side" factors at play. If employers think that obese workers are likely to be less productive or likely to be more expensive to employ because of health care costs, then obese workers may have a more difficult time finding a job than similarly qualified workers who are not obese.

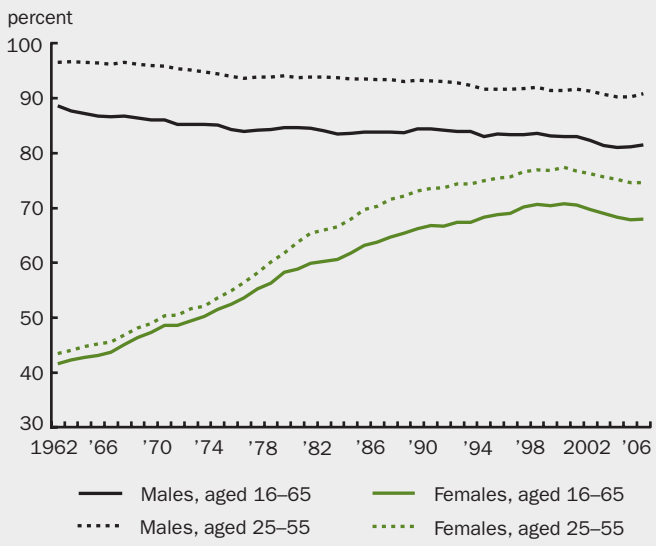
In this article, we examine both self-reported health and disability outcomes and employment outcomes to try to distinguish between supply side and demand side explanations. If, for example, there is no change in the relationship between obesity and health outcomes, but there is a change in the relationship between obesity and employment outcomes, that would suggest that demand side factors might play an important role in non-employment among the obese.

We are also interested in whether the changes we observe over time in health and employment outcomes are due to changes in the underlying population characteristics, such as a rising incidence of obesity, or due to an increase in the differences in outcomes between the obese and the nonobese. For example, if in every period the obese are more likely to be in poor health than the nonobese, then an increase in the proportion of the population that is obese will likely lead to a larger proportion of the population that does not work. On the other hand, the *propensity* to report poor health, disability, or non-employment among the obese compared with the nonobese may also have changed over time. This change in propensities may be due to either supply side or demand side factors that are shaped by changes in health policies and/or labor market policies. For example, in 1984 there was a substantial change in disability insurance (SSDI) criteria that may have made it more likely that someone with obesity-related

Kristin F. Butcher is an associate professor of economics at Wellesley College and a former senior economist at the Federal Reserve Bank of Chicago. Kyung H. Park is a senior associate economist at the Federal Reserve Bank of Chicago. *The authors thank Dan Sullivan, Anna Paulson, Bhashkar Mazumder, and seminar participants at the Federal Reserve Bank of Chicago for helpful comments. The views expressed here are those of the authors and do not necessarily represent those of Wellesley College or any other entity.*

FIGURE 1

Labor force participation, by gender and age group



Source: Authors' calculations based on data from the U.S. Census Bureau, March *Current Population Surveys*.

health conditions could qualify for SSDI. This change, combined with subsequent changes in the wage structure that made SSDI benefits more generous relative to low-wage jobs, may have made some obese people more likely to opt out of the labor market. Thus, an increase in the number of obese people in the population would have a different effect on outcomes, depending on the period in which the change is evaluated.

We find that, although those who are heavier have always had worse self-reported health outcomes and employment outcomes, there is not much evidence that the propensity for the obese to have poor outcomes has changed over time. Non-employment among men of prime age increased from 10 percent in 1984–85 to 12.5 percent in 2004–05. Increases in obesity alone can explain about 3 percent to 12 percent of that increase. In addition, population changes in age, race, and ethnicity, combined with changes in obesity, can explain between 34 percent and 47 percent of the increase in men's non-employment. These results suggest that deterioration in underlying health has played an important role in the decrease in men's labor force participation and that these population changes would have had similar effects whether evaluated in the mid-1980s or early 2000s.

In the next section, we describe recent trends in non-employment and labor force participation, age, obesity, and disability insurance receipt. We examine whether the propensity for the morbidly obese to self-report musculoskeletal conditions and routine needs disability (defined as requiring the assistance of another person in handling routine tasks, such as personal care, housework, or shopping) and to apply for disability insurance has changed over time. Then, we analyze how much of the change in non-employment can be explained by changes in obesity and other demographic characteristics.

Changes in non-employment, age, obesity, and disability insurance

First, we look at the changes in labor force participation by gender and age group from 1962 through 2006, using the March *Current Population Survey* (CPS), which is conducted by the U.S. Census Bureau for the U.S. Bureau of Labor Statistics (figure 1). Clearly, labor force participation among women rose dramatically from the 1960s through the 1990s and leveled off in the 2000s. The change has been less dramatic for men, but over the same period, we have seen a continuous decline in men's labor force participation. Note that this is the case even for relatively young men (aged 25–55).

If we look at the share of survey respondents who reported that they had not worked the previous week (we call this the share "not working last week")—which includes nonparticipants and the unemployed—we see a similar pattern (figure 2). While the share not working has declined for women, it has risen for men. Again, this is true even among relatively young men.

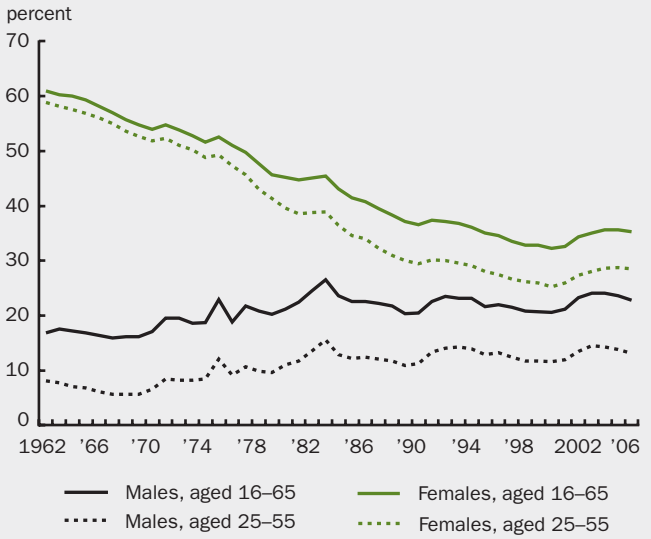
Changes in the age distribution

Some of the changes in the labor supply documented in the previous section may be related to changes in the age distribution. Figure 3 shows the shift in the age distribution among all 25–54 year olds between 1976–80 and 1999–2002. As the baby boom generation ages, there is a change in the average age among 25–54 year olds. For women, labor supply peaks prior to childbearing and again once their children are older. For men, Barrow and Butcher (2004)[2] show that in both 1978–79 and 1999–2000 periods, the fraction of men who did not work at all in the previous year increased monotonically across age groups for those above age 40. Since morbidity increases with age, it seems likely that the aging of the population—even among men aged 25–54—would lead to increases in non-employment.

Barrow and Butcher (2004) point out that there have been other demographic changes, for example,

FIGURE 2

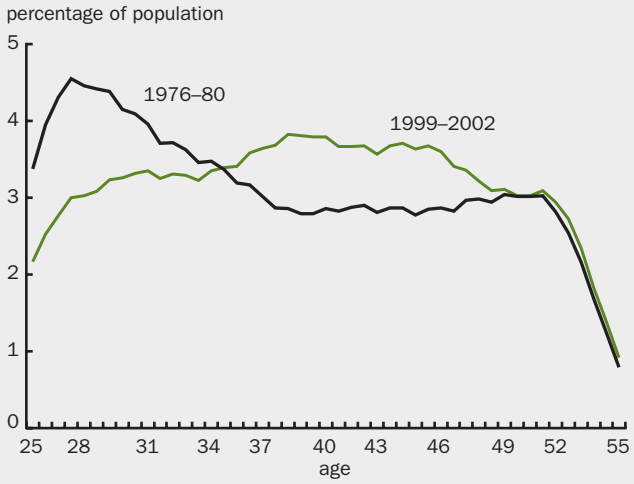
Share not working last week, by gender and age group



Source: Authors' calculations based on data from the U.S. Census Bureau, March *Current Population Surveys*.

FIGURE 3

Age distribution



Note: The distribution is among those aged 25–54.

Source: Authors' calculations based on data from the U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, *National Health and Nutrition Examination Survey*.

changes in the racial and ethnic mix of the population, that may also be correlated with deteriorating health. Their analysis, which does not control for obesity, finds that 14 percent to 33 percent of the increase in men's full-year non-employment that occurred between 1978–79 and 1999–2000 can be attributed to changes in age, race, and ethnicity alone.

Changes in obesity

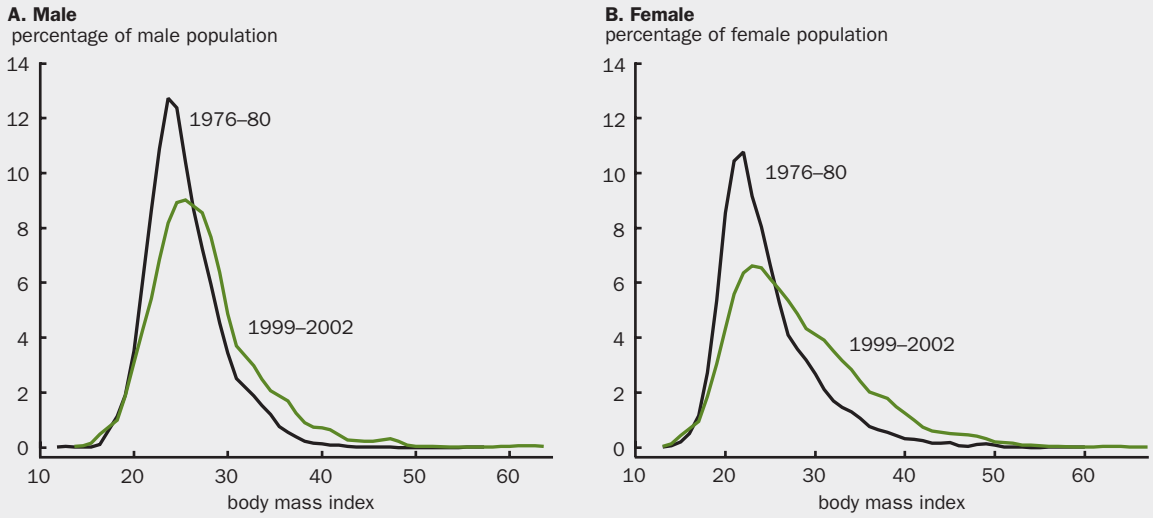
Although many of the demographic changes over the past 30 years might lead us to expect a deterioration of health in the working age population, many health indicators suggest improvements in health or improvements in individuals' quality of life, even when they have a health problem (Cutler and Richardson, 1997). However, obesity has become increasingly common during this period. Obesity is typically defined using the body mass index.[3] A BMI lower than 18.5 is considered underweight; a BMI lower than 25 (but not lower than 18.5) is considered a healthy or normal weight; a BMI greater than or equal to 25 is deemed overweight; a BMI greater than or equal to 30 is deemed obese; and a BMI greater than or equal to 40 is considered morbidly obese.

Figure 4 shows the probability density function for BMI for men and women aged 25–54 years old in the 1976–80 and 1999–2002 *National Health and Nutrition Examination Surveys* (NHANES), which are conducted by the U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. These distributions show the rightward shift in the BMI distribution over time.

Although there has been an increase in median BMI, a significant feature underlying the obesity epidemic is that the variance in BMI has increased. The heavy have gotten much heavier over time. Panels A and B of figure 5 highlight these changes in the distribution of BMI, using NHANES data for men and women, respectively. Note the median male in 1999–2002 would have been heavier than nearly three-quarters of the population in the earlier period 1976–80. A male just on the cusp of obesity (75th percentile) in the 1999–2002 BMI distribution would have been heavier than 90 percent of the earlier period's population. For females, we also see dramatic changes in the BMI distribution in the heaviest portions of the distribution.

FIGURE 4

Body mass index density

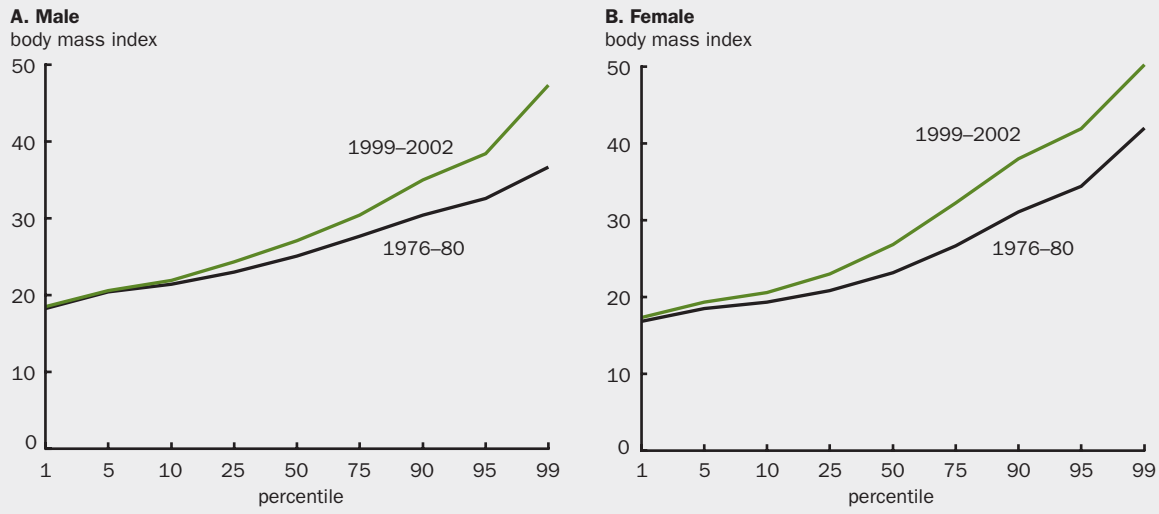


Notes: The density is calculated for those aged 25–54. A body mass index lower than 18.5 is considered underweight; 18.5–24.9, normal weight; 25.0–29.9, overweight; 30.0–39.9, obese; and 40.0 or higher, morbidly obese.

Source: Authors' calculations based on data from the U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, *National Health and Nutrition Examination Survey*.

FIGURE 5

Body mass index distribution



Note: The distribution is among those aged 25–54.

Source: Authors' calculations based on data from the U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, *National Health and Nutrition Examination Survey*.

If it is the very heavy who are most likely to suffer ill health from obesity, then the population at risk of obesity-related health conditions has increased. Further, if being heavy is more likely to cause one health problems as one ages, then we would expect that as these heavier cohorts age, they will experience more weight-related health problems than previous, slimmer cohorts.

Figures 1 through 5 demonstrate that non-employment among men of prime age has increased.

They also document shifts in the population—namely, the population is older and more likely to be obese—that are consistent with a health-based reason for this decline in work among men.

Changes in disability insurance

Figure 6 shows that the percentage of the population receiving disability benefits has risen substantially since the early 1980s and that the increase seems to have begun after 1984.[4] Changes to the disability insurance eligibility rules in 1984 appear to have increased the likelihood that an SSDI applicant would receive payments. As Autor and Duggan (2003) explain, the awards criteria now give more weight to an individual's pain and ability to function in the work place; prior to 1984, eligibility was determined by "continuous disability reviews" by third party physicians. In addition, rising wage inequality during the 1980s and 1990s increased the value of SSDI payments relative to wages for many individuals. Many observers have linked these changes in the SSDI program to increases in disability insurance receipt and decreases in employment.

Coinciding with these programmatic changes, there have been changes in the primary diagnoses among recipients. Table 1 documents the share of disability awards attributed to different disorders. In 1981, prior to the new disability insurance eligibility criteria, 17 percent of all awards were for musculoskeletal disorders; by 2003, this figure had risen to 26.3 percent. Mental disorders have also accounted for an increased share of SSDI awards since 1981.

Figure 7 demonstrates how SSDI awards for various causes have changed on a population basis (per 10,000 individuals, aged 16–64). Heart disease and cancer have held steady as reasons for disability insurance claims, but musculoskeletal conditions, mental illness, and other sources have increased.[5] This shift in the reasons documented for disability receipt is often seen as being due to changes in the criteria used to judge whether an individual is disabled. Diseases that are easily verifiable by a physician—for example, cancer and heart disease—have declined as a share of all disability awards. This is not to say, however, that there have not also been changes in underlying health that would contribute to these shifts in disability insurance payments.

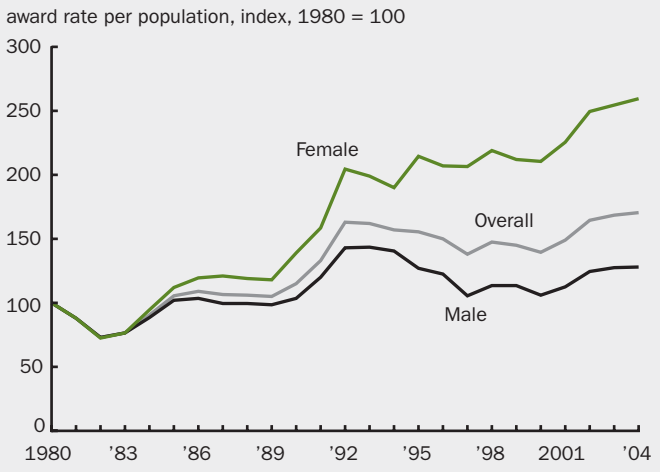
For example, there are many ways that the increase in obesity may be related to the increase in the share of disability awards for musculoskeletal disorders. It may be that the increase in obesity has led to more musculoskeletal disorders, in turn leading to more disability claims. In this case, the driver of the increase is the change in obesity rates leading to more musculoskeletal disorders. On the other hand, changes in disability insurance rules—which now give more emphasis to an individual's report of pain—may have also given those who are obese, and thus have a better basis for making a claim of musculoskeletal pain, a better chance to qualify for SSDI. Changes in wages relative to SSDI payments may have given workers an increased incentive to apply for disability insurance.

In the next section, we examine whether the propensity of the obese to claim various health ailments, to self-report routine needs disability, or to apply for SSDI has changed over time. The 1984 change in the SSDI rules does not fall in the span of our data on self-reported health, so this exercise does not shed light on how that policy change may have affected behavior. Instead it allows us to answer the following question: During the period after 1984 when awards for musculoskeletal disorders continue to rise, do we see a rise in the propensity of the obese to report these ailments?

In the rest of this article, we focus only on men aged 25–54 years old, since it is this group that has shown a rising

FIGURE 6

Social Security Disability Insurance award rate per population



Note: The sample population is made up of those aged 30–54.
Sources: Authors' calculations based on data from the U.S. Social Security Administration, *Annual Statistical Supplement to the Social Security Bulletin, 2005*; and Haver Analytics.

TABLE 1

Share of total Social Security Disability Insurance awards, by diagnosis

Diagnosis	Total 1981	Total 2003	Males 1981	Males 2003	Females 1981	Females 2003
	(percentage)					
Heart disease	24.9	11.4	27.9	14.4	18.0	7.8
Cancer	16.3	9.4	15.1	9.1	19.1	9.7
Mental disorders	10.5	25.4	10.2	23.0	11.2	28.2
Musculoskeletal disorders	17.0	26.3	15.6	24.7	20.1	28.2
Nervous system	8.3	8.5	7.9	8.2	9.1	8.9
Respiratory system	6.2	4.2	6.5	4.1	5.5	4.4
Endocrine system	4.3	3.1	3.8	3.1	5.5	3.1
All other	12.5	11.6	12.9	13.5	11.5	9.9
Total	100.0	99.9	100.0	100.1	100.0	100.2

Note: Columns may not total because of rounding.
Sources: Authors' calculations based on data from the U.S. Social Security Administration, *Annual Statistical Supplement to the Social Security Bulletin, 1981*, and *Annual Statistical Report on the Social Security Disability Insurance Program, 2003*.

trend in non-employment over this period. The underlying health conditions have changed in similar ways for women, making their large increase in labor force participation even more striking.

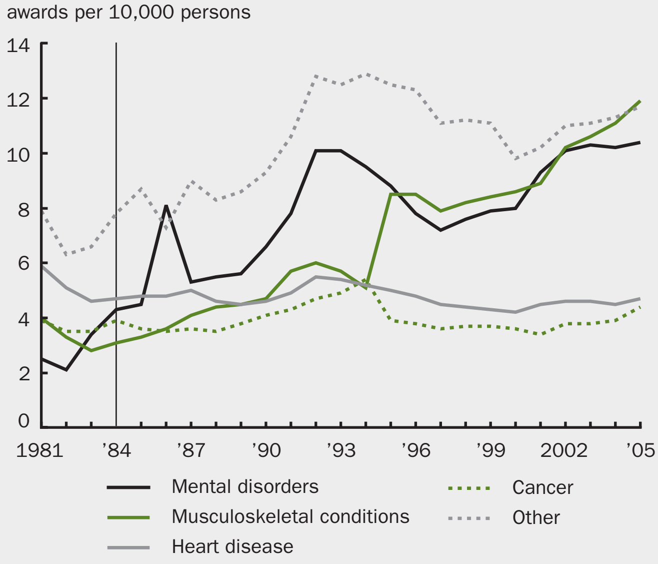
Self-reported health conditions, disability, SSDI receipt, and obesity

Since one can report a health condition without claiming to be disabled by it and since one can claim to have a disability without applying for disability insurance, we examine the relationship between obesity and each of these outcomes separately. We show how the relationship has changed over time. We are particularly interested in whether the propensity for those who are heavy to report poor health outcomes has increased over time, which would be consistent with changes in the incentives of the obese to apply for SSDI and leave the labor force.

Figure 8 shows the unadjusted prevalence of musculoskeletal disorders for men who are underweight, normal weight, overweight, obese, and morbidly obese**.** From 1984 through 1996, those who are heavier are more likely to report a musculoskeletal problem. There is an increase in reports of musculoskeletal problems among the morbidly obese from 1984 through 1988, but there is a decline in later years. In general, there is little evidence of an increase in the propensity for the obese and morbidly obese to report a musculoskeletal problem. This finding may be somewhat misleading, however, because it does not control for other demographic differences that may be correlated with obesity and with reports of musculoskeletal problems. To address this, we use regression

FIGURE 7

Social Security Disability Insurance awards per population, by diagnosis



Notes: The sample population is made up of those aged 16–64. Changes to the disability insurance eligibility rules were made in 1984. See the text for further details.
Source: Authors' calculations based on data from the U.S. Social Security Administration, *Annual Statistical Report on the Social Security Disability Insurance Program, 2005*.

analysis, which allows us to hold constant other demographic differences and examine whether the likelihood of reporting a given health issue has changed over time by weight category.

The *National Health Interview Survey* (NHIS)—conducted by the U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics—asks a series of health questions that allow us to examine components of musculoskeletal disorders. Figure 9 presents differences in reporting of lower back pain between morbidly obese men and those of normal weight in the 1997–2005 *National Health Interview Surveys*. We calculated these differences by running a linear probability model on whether the individual reports lower back pain, controlling for indicator variables for underweight, overweight, obese, and morbidly obese. Normal weight is the omitted category. Only the morbidly obese were statistically significantly more likely to report these ailments. We ran separate regressions without any controls, as well as controlling for age alone and then controlling for age, race, and Hispanic ethnicity.[6] We ran a separate regression for each year, thus allowing the effect of the regressors to differ each year. (Figures 9, 10, and 11 also include the 95 percent confidence intervals for the difference in reporting between the morbidly obese and those of normal weight.)

We see that over this period, those who are morbidly obese are more likely to report lower back pain, although for some years this difference is not statistically significantly different from zero. Although the point estimate for the difference in reporting lower back pain is higher later in the period, the difference in the effects between the two periods is not statistically significant. Thus, there is little evidence of an increase in the difference in reports of lower back pain between the morbidly obese and those who are of normal weight during this period. Also, note that our estimates do not vary substantially as we add control variables. The results are similar for

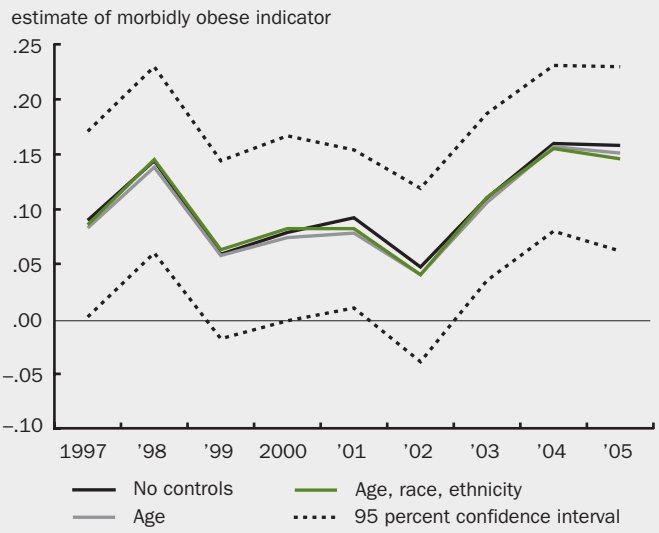
FIGURE 8

Prevalence rates of chronic musculoskeletal conditions among males

Note: The sample population is made up of males aged 25–54.
Source: Authors' calculations based on data from the U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, *National Health Interview Survey*.

FIGURE 9

Incidence of lower back pain: Morbidly obese vs. normal weight males



Note: The sample population is made up of males aged 25–54.
Source: Authors' calculations based on data from the U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, *National Health Interview Survey*.

other components of musculoskeletal disorders, such as reported arthritis or other joint pain.

Figure 10 examines whether the morbidly obese have become relatively more likely over time to report routine needs disabilities. The data are from the NHIS from 1984 through 2005. There were significant changes in sequence and wording of the disability questions between 1996 and 1997, and thus, we show a break in the series.[7] The figures are based on linear probability models that are analogous to those described for figure 9.

Again, we see that the morbidly obese are more likely to report a routine needs disability; and controlling for age, race, and ethnicity makes little difference in the size of that effect. However, there is no statistically significant difference in the size of the effect of morbid obesity across time periods.

Finally, figure 11 shows the difference in the probability of ever having applied for disability insurance between the morbidly obese and those categorized as having normal weight, controlling for age, race, and ethnicity. Information on applications for disability insurance are only available after 1996, and all respondents are asked if they have "ever applied for" disability insurance. While the morbidly obese have always been statistically significantly more likely to have applied for disability insurance than those of normal weight, this difference is stable over the period observed.

Table 1 (p. 7) and figures 6 (p. 6) and 7 (p. 7) in the previous section showed that disability awards have been increasing since the mid-1980s, particularly for musculoskeletal ailments. In this section, we examined the relationship between obesity, health, disability, and application for SSDI. The evidence shows that obesity has increased, with morbid obesity having increased in particular. In addition, since the mid-1980s the morbidly obese, in particular, have reported worse health outcomes than their nonobese counterparts. However, over this period we have not seen an increase in the propensity to report worse health outcomes by the morbidly obese, nor an increase in the likelihood of their applying for SSDI. What we have seen is that there are now more of the category of people—very obese people—who have always reported worse health outcomes, but not much evidence of an increase in the likelihood of reporting worse health outcomes among the very obese.

Non-employment and obesity

In this section, we examine the relationship between obesity and employment. This relationship may be different from the relationship between obesity

FIGURE 10

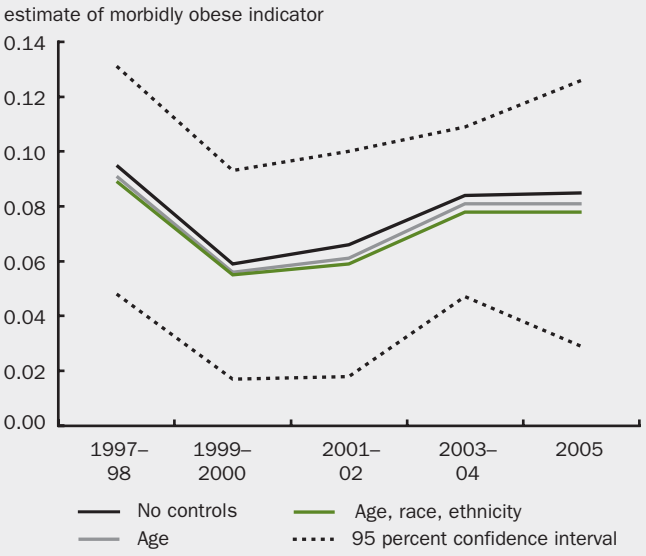
Routine needs disability: Morbidly obese vs. normal weight males

Notes: The sample population is made up of males aged 25–54. Routine needs disability is defined as requiring the assistance of another person in handling routine tasks, such as personal care, housework, or shopping. There were significant changes in sequence and wording of the disability questions between 1996 and 1997, and thus, we show a break in the series. See note 7 for further details.

Source: Authors' calculations based on data from the U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, *National Health Interview Survey*.

FIGURE 11

Ever applied for Social Security Disability Insurance: Morbidly obese vs. normal weight males



Note: The sample population is made up of males aged 25–54.
Source: Authors' calculations based on data from the U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, *National Health Interview Survey*.

and self-reported health measures. For example, across a number of periods, an increase in obesity may affect health in a similar way. However, the employment response to that change in health may differ, depending on both demand side (from the employers' perspective) and supply side (from the workers' perspective) changes in the employment–obesity relationship.

First, there is some debate about the ways in which changes in employment itself may have contributed to the rise in obesity (Philipson and Posner, 1999). For many of us, technological changes have tended to reduce the calories we expend at work by letting us spend more time at our desks. This is true even in employment sectors that typically required more physical activity, as more and more processes in industrial and manufacturing environments have become automated. This trend may have contributed to the long-term increase in BMI, although much of the recent rise in obesity seems to have begun in the 1980s, when one might argue that the transition from hard physical labor to sedentary work had already happened. Nonetheless, that transition may have important implications for the effect of obesity on one's ability to work—if most people at work are engaged in sedentary tasks that require little physical exertion, then the effect of obesity on the ability to perform a job may be smaller in the current technological era than it would have been when heavy physical exertion was a frequent requirement at work.

In addition, there is some evidence of discrimination against obese people (see Carpenter, 2006, and Cawley and Danziger, 2004). Suppose there are two equally productive individuals—one obese and one not obese—and employers are less willing to hire the obese individual. If that preference for the nonobese was constant over time, the increase in the obese population could lead to an increase in the fraction of individuals who are not working. In addition, however, employers' "preference" for hiring nonobese people could change over time. On the one hand, technological changes that reduce the physical requirements of jobs would seem to narrow any perceived productivity gap between obese and nonobese workers. On the other hand, even if productivity is not a concern, the rising costs of employer-provided health insurance may make employers less inclined to hire those they perceive as being costly employees over time.

Finally, of course, those who are obese may be less likely to work than individuals of normal weight for other, more personal reasons. They may be in poorer health, making work more difficult, or they may find work less enjoyable than their counterparts of normal weight. Changes in working conditions may also have an impact on obese workers. These conditions could include demand side factors, discussed previously, or supply side factors. If, for example, wages for the obese fall or SSDI becomes either easier to get or more generous relative to the wages they could likely command, then the obese might change their propensity to work in a given period.

In the analysis that follows, we want to disentangle the increase in non-employment that has arisen because there are more obese people, and particularly more *morbidly* obese people, today than there were 20 years ago from any increase that has occurred because the effect of obesity on non-employment has changed.[8]

We focus on measures of non-employment that are available in the data sets that also track obesity over time. The two main data sets are the *National Health Interview Survey* and the *National Health and Nutrition Examination Survey*. Note that the information available in the data set usually used to track labor

market statistics (CPS) and the information available in the data sets usually used to track health statistics (NHIS and NHANES) are not the same. In particular, the data sets that contain information on BMI and obesity have less detailed information on whether one is working. In the CPS, one can examine the fraction of the year spent not working, for example, or the fraction of the population that is not employed for the entire year (see Barrow and Butcher, 2004). In the health data sets, the available data restrict us to classifying people as non-employed if they report not working in the previous one to two weeks. Table 2 compares the health and labor force data available in the *Current Population Surveys*, *National Health Interview Surveys*, and *National Health and Nutrition Examination Surveys*.

Table 3 shows the differences in the reported share of non-employed by year using the different data sets. We see that the NHIS closely tracks the non-employment figures calculated from the CPS. In contrast, the NHANES overstates the growth in non-employment among men in the prime age category by more than twofold. For this reason, we focus on the NHIS in the analysis that follows.

In order to examine how much of the change in non-employment can be explained by changes in obesity, we use an Oaxaca–Blinder multivariate decomposition (see Oaxaca, 1973, and Blinder, 1973). Here, we run linear probability regressions with not working in the past one to two weeks as the outcome variable. We control for underweight, overweight, obese, and morbidly obese as the weight categories, with the normal weight category omitted. In some regressions, we also control for age, race, and ethnicity, as well as for pairwise interactions between weight categories and age and race. We run these regressions in both the early (1984–85) and later (2004–05) years of our data series:

$$1) \quad Y_{84-85} = \beta^0_{84-85} + \beta^1_{84-85} X_{84-85} + \varepsilon_{84-85};$$

$$2) \quad Y_{04-05} = \beta^0_{04-05} + \beta^1_{04-05} X_{04-05} + \varepsilon_{04-05}.$$

Typically, these equations are then rearranged to examine how much of the difference in outcomes between the two years is due to differences in the explanatory (X) variables, and how much is due to differences in the effects of these variables on the outcomes, the β values. Differences attributable to changes in the

TABLE 2

Comparison of labor force and health data, by data source

Labor force data	CPS March	NHIS 1984–96	NHIS 1997–2005	NHANES 1976–80	NHANES 1999–2002
Worked last 1–2 weeks	X	X	X	X	X
Reason not working last week	X		X	X	X
Class of worker	X	X	X	X	X
Hours worked last week	X		X		X
Full/part time	X		X	X	X
Weeks worked	X				
Months worked	X		X		X
Wage data	X				
Industry	X	X	X	X	X
Occupation	X	X	X	X	X
Health data					
Body mass index or weight/height		X	X	X	X
Disability/physical limitations	X	X	X	X	X
Conditions causing disability		X	X	X	X
Ever applied for Social Security Disability Insurance		X	X		

Notes: CPS means *Current Population Survey*. NHIS means *National Health Interview Survey*. NHANES means *National Health and Nutrition Examination Survey*. In the NHIS 1984–96 and NHANES 1976–80, the employment status question asks whether or not the respondent has worked in the past two weeks, while the NHIS 1997–2005 and NHANES 1999–2002 ask about employment status in the past one week. The March CPS employment status variables (*esr* and *mlr*) also ask about employment status in the past one week. The March CPS also asks questions related to disability status. One variable notes whether or not "health or disability limits kind or amount of work." Another records whether someone left a job for health reasons. Finally, the data include a variable indicating whether or not the household receives disability income.

Sources: U.S. Census Bureau, March *Current Population Surveys*; and U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, *National Health Interview Survey* and *National Health and Nutrition Examination Survey*.

TABLE 3

Comparison of share of non-employed males, by data source

	CPS	NHIS
	(- - - - - - - - percentage - - - - - - -)	
2004–05	13.4	12.5
1984–85	11.5	10.3
Change	2.0	2.2
	CPS	**NHANES**
1999–2002	11.9	12.0
1976–80	9.8	7.5
Change	2.1	4.6

Notes: The sample population is made up of males aged 25–54. Columns may not total because of rounding. CPS means *Current Population Survey*. NHIS means *National Health Interview Survey*. NHANES means *National Health and Nutrition Examination Survey*.

Sources: Authors' calculations based on data from the U.S. Census Bureau, March *Current Population Surveys*; and U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, *National Health Interview Survey* and *National Health and Nutrition Examination Survey*.

X variables are attributable to changes in obesity, age, race, and ethnicity.[9] Differences attributable to changes in the coefficients, on the other hand, are attributable to the supply side and demand side factors described previously.

$$3)\quad Y_{04-05} - Y_{84-85} = \beta^1_{04-05}(X_{04-05} - X_{84-85}) + (\beta^1_{04-05} - \beta^1_{84-85})X_{84-85} + (\beta^0_{04-05} - \beta^0_{84-85}).$$

The first term after the equals sign is the difference attributable to changes in the X values, and the second two terms are the differences attributable to changes in the coefficients. As written out in equation 3, the change in individual characteristics between the two periods is evaluated using the "returns" to these characteristics that prevailed in the later period. If we had done the subtraction the other way, we would get a different answer.

Our approach is to examine how the changes in individual characteristics that actually occurred between 1984–85 and 2004–05 would have been expected to change the fraction of the population that was not working, given the "conditions" that prevailed in both the earlier and later periods. We can use equations 1 and 2 to predict how people with the characteristics of those who existed in 2004–05 would have "behaved" in 1984–85:

$$\beta^1_{84-85}X_{04-05}.$$

And we can use those same equations to predict how people with the characteristics of those who existed in 1984–85 would have "behaved" in 2004–05:

$$\beta^1_{04-05}X_{84-85}.$$

Suppose we imagine that the only thing that explains the increase in men's non-employment is that non-employment is higher among the morbidly obese and that, in the later period, more men are morbidly obese. Then, evaluating the effect of the increase in morbid obesity using the "returns" to morbid obesity that prevailed in the earlier period should yield the exact increase in non-employment that we observe in the data. Since, in fact, conditions, or "returns to characteristics," may have changed, we can think of this exercise as answering the following question: How much of an increase in non-employment would we have expected in 1984–85 if morbid obesity had increased to today's levels under those conditions?

We present these calculations, allowing age, race, and ethnicity characteristics to change in addition to obesity measures, and we allow for pairwise interactions in these characteristics. Age may exacerbate the health problems associated with obesity—for example, the knees of 30 year olds may not hurt among either those of normal weight or the obese, but the knees of 50 year olds may have suffered more wear and tear among the obese but still be fairly pain free among those of normal weight. And thus, we would find that adjusting the data from the two periods to have the same age–obesity profile explains more of the change in non-employment over time. Obesity may have different effects in different populations as well as for different age groups. If obesity-related health problems are more prevalent among blacks and Hispanics, for example, then adjusting for the obesity–age–race/ethnicity profile may explain more of the changes over time. We present the results for these different adjustments separately.

Our decompositions are similar to those presented in Lakdawalla, Bhattacharya, and Goldman (2004). They examine how much of the increase in *disability rates* across different age groups between 1984 and 1996 can be explained by the rise in obesity. They decompose the change in disability rates between 1984 and 1996 into:

$$[(O_{96} - O_{84}) * (D^O_{90} - D^{NO}_{90})] + [O_{90} * \{(D^O_{96} - D^O_{84}) - (D^{NO}_{96} - D^{NO}_{84})\}],$$

where O_{yr} is the obesity rate in a given year and D_{yr} is the disability rate in a given year, and where the superscripts denote whether the disability rate is measured among the obese (*O*) or the nonobese (*NO*).

The first term in this expression is the amount of increased disability we would have expected had obesity risen as it did between the two periods, but the effect of obesity on disability was as it was in the interim year—1990. The second term is the amount of increase in disability that is due to the fact that disability among the obese rose, holding constant obesity rates at the level of the interim period. Using this decomposition, Lakdawalla, Bhattacharya, and Goldman (2004) find that 50 percent of the rise in disability for 18–29 year olds; 25 percent for 30–39 year olds; 10 percent for 40–49 year olds; and nearly all for 50–59 year olds can be explained by increases in obesity.[10]

This calculation combines the rise in disability that comes from the increase in obesity and the rise in disability that comes from changes in the *effect* of obesity on disability. In our analysis that follows, we focus on numbers that are similar to the first component—the amount by which *non-employment* would have risen had obesity rates risen—but we show this effect under the conditions of the earlier and later periods—that is, holding constant the effect of obesity on non-employment at its level in the earlier period and then at its level in the later period.

Table 4 presents the results of these simulations. The first row shows actual non-employment rates, which increased 2.2 percentage points, from 10.3 percent to 12.5 percent between 1984–85 and 2004–05. The second row shows predicted non-employment rates given the BMI distribution that existed in the other period, using the coefficients for the period listed in the column heading. For example, looking at the second row of numbers, the first column tells us that had the weight distribution that existed in 2004–05 occurred in 1984–85, we would have seen a non-employment rate of 10.4 percent in 1984–85—slightly higher than the actual non-employment rate in that period. Similarly, if the weight distribution that existed in 1984–85 occurred in 2004–05, we would expect a non-employment rate of 12.3 percent—slightly lower than the actual non-employment rate in that period. The last two columns show us how much of the actual change in non-employment between the two periods can be explained by evaluating the change in characteristics listed on the leftmost column using the returns to those characteristics in the years given in the column headings. So, about 3 percent of the increase in non-employment can be explained by the rise in obesity alone using the "returns" to obesity that prevailed in 1984–85. About 13 percent of the rise in non-employment would be attributed to the increase in obesity if we evaluated that increase using the "returns" that prevailed in 2004–05.[11] This is consistent with a story in which either supply side or demand side deterrents to working for the obese are stronger in 2004–05 than in 1984–85. For example, this could occur if disability insurance takeup rates are higher among the obese in the later period. However, if there are other characteristics of obese workers that are also correlated with non-employment but are not held constant in these regressions, then those effects will load onto the obesity coefficients here, leading us to attribute either too little or too much of the changes to changes in obesity. Furthermore, changes in the characteristics we use in our analysis—age, race, and ethnicity—may also be linked to changes in underlying health. Finally, as discussed earlier, we want to include interactions between age, race, ethnicity, and weight

TABLE 4

Actual and simulated average share of non-employed males and the percent of actual change explained by given characteristics

			Percent of actual increase explained by characteristics under conditions in:	
	1984–85	2004–05	1984–85	2004–05
	(percentage)			
Actual non-employment	10.3	12.5		
Characteristics used in simulation				
Weight categories	10.4	12.3	3.4	12.5
Weight categories, age polynomial	10.7	11.8	14.3	31.6
Weight categories, age, race, ethnicity (all interactions)	11.4	11.8	46.8	33.9

Notes: The sample population is made up of males aged 25–54. The normal weight category is excluded from the weight categories. See the text for further details.

Source: Authors' calculations based on data from the U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, *National Health Interview Survey*.

measures. If it is not just the fraction of the population that is morbidly obese that matters for non-employment, but rather the fraction that is older and morbidly obese, we want to capture that in our simulations.

Including a polynomial in age in our simulations increases the amount of the increase in non-employment that we can explain to 14 percent using the earlier period and 32 percent using the later period. Once we include age, race, and ethnicity in the models, more of the increase in non-employment can be explained using the returns to characteristics that prevailed in 1984–85 than those in 2004–05. Changes in these characteristics can explain from 34 percent (using 2004–05 returns to characteristics) to 47 percent (using 1984–85 returns to characteristics) of the increase in the non-employment rate.[12]

Changes in age, race, and ethnicity—which may themselves be markers of changes in underlying health—explain a larger share of the increase in non-employment between 1984–85 and 2004–05 than do changes in obesity measures alone. However, (in results not shown) adding obesity measures to simulations that include age, race, and ethnicity controls increases the amount of the predicted increase in non-employment by 10 percentage points, regardless of which period we use to evaluate the change.

These results suggest that changes in underlying population characteristics may have played an important role in the increase in non-employment among men of prime working age over the past 30 years.

Conclusion

This article examines the role of the increase in obesity in changes in non-employment. Men of prime working age have increased their non-employment rates over the past 30 years, and disability rates have also increased. Many have noted that this increase has happened against a backdrop of generally improving health in the U.S. population. However, obesity has increased substantially over this period. Here, we have tried to disentangle the changes that occurred in heath and employment because of the increase in the fraction of the population that is obese from the changes that are due to changes in the differences in outcomes between obese and nonobese individuals. We find that, while the morbidly obese have always been more likely to report musculoskeletal ailments and more likely to report being disabled, their propensity to report ailments and disability has not statistically significantly increased over time.

The results for non-employment are consistent with those for health and disability. If the results had shown that increases in obesity had little effect on health and disability rates but had a large effect on employment, this would have pointed toward the importance of demand side factors—such as efforts by employers to avoid higher health care costs—in employment outcomes for the obese. However, since the results are consistent for health, disability, and non-employment, we cannot use these differences to infer the relative importance of demand side or supply side effects.

For men of prime working age, changes in their characteristics—including age, race, ethnicity, and obesity levels—can explain a large portion (around 40 percent) of the increase in non-employment over the period. The portion of the change in non-employment that is explained by changes in these characteristics is similar regardless of whether we evaluate the change in characteristics using the returns to characteristics that prevailed in either the earlier period (1984–85) or the later period (2004–05). This means that under either the earlier or later labor market conditions, we would expect that these changes in characteristics would lead to a substantial increase in non-employment. Similar to Lakdawalla, Bhattacharya, and Goldman (2004), we find that the obesity epidemic may be playing an important role in changing labor market outcomes.

NOTES

[1]See Barrow (2004); Anderson, Barrow, and Butcher (2005); and Aaronson, Park, and Sullivan (2006) for trends in unemployment rates and labor force participation.

[2]A 2006 revision to Barrow and Butcher (2004) is available from the authors upon request.

[3]Body mass index = (weight in kilograms)/(height in meters squared).

[4]Disability benefit award numbers are from the Social Security Administration's (SSA) *Annual Statistical Supplement to the Social Security Bulletin, 2005*, and the noncivilian population figures come from the monthly household data in Haver Analytics. Note that disability awards have risen particularly among women. Disability insurance pays benefits to an individual and certain family members, provided that the individual is "insured"—meaning that the person has worked long enough and paid social security taxes. The increase in women's labor supply presumably increased the pool of eligible workers. See www.ssa.gov/disability/.

[5]The substantial increase in musculoskeletal conditions in 1995 is due to a different sampling methodology used by the Social Security Administration. Prior to 1995, the SSA only included awards allowed after the initial determination. Since many musculoskeletal conditions are denied initially and awarded later after an appeals process, the pre-1995 sample understates the share of musculoskeletal awards relative to the post-1995 sample that includes awards granted after the appeals process.

[6]Specifications include age and age squared, as well as indicator variables for black, other, and Hispanic ethnicity.

[7]Prior to 1997, only respondents who had a major activity limitation were asked if they needed assistance with personal care or routine need tasks. Individuals older than 60 years, however, were not screened and were automatically asked about any potential disability. In 1997, respondents were no longer screened and everyone was asked about personal care or routine needs disability. In 1997, the wording of the disability question also changed. Previously the personal care question read, "Because of any impairment or health problem, does ___ need the help of other persons with personal care needs, such as eating, bathing, dressing, or getting around this home?" After 1996, however, the question read, "Because of a mental, physical, or emotional problem, does ____ need the help of other persons with personal care needs, such as eating, bathing, dressing, or getting around this home?"

[8]The former will be changes in the characteristics of the population (the Xs) and the latter changes in coefficients (the βs).

[9]Specifically, weight categories (underweight, overweight, obese, and morbidly obese), age, age squared, black, other, Hispanic ethnicity, and interactions between the weight categories and the other demographic variables (age and race) are included in the X values.

[10]Their analysis also includes women.

[11]This is because the point estimate for the coefficient on morbid obesity is higher in the later years; however, just as in the results for health conditions presented earlier, this difference is not statistically significant.

[12]We find similar results if we decompose the change in routine needs disabilities. Because of the change in survey questions regarding routine needs disabilities, we perform this analysis for changes from 1984–85 through 1995–96 and from 1996–97 through 2004–05. Changes in weight categories, age, race, and ethnicity can explain about a third of the increase in routine needs disabilities between 1984–85 and 1995–96, using the "returns" to these characteristics that prevailed in either time period. Changes in these characteristics between 1996–97 and 2004–05 explain about 33 percent of the increase in routine needs disabilities using the "returns" to characteristics that prevailed in 1996–97 and about 42 percent of the increase using "returns" that prevailed in 2004–05.

REFERENCES

Aaronson, Daniel, Kyung-Hong Park, and Daniel Sullivan, 2006, "The decline in teen labor force participation," *Economic Perspectives*, Federal Reserve Bank of Chicago, Vol. 30, No. 1, First Quarter, pp. 2–18.

Anderson, Katharine, Lisa Barrow, and Kristin F. Butcher, 2005, "Implications of changes in men's and women's labor force participation for real compensation growth and inflation," *Topics in Economic Analysis and Policy*, Vol. 5, No. 1, article 7.

Autor, David H., and Mark G. Duggan, 2003, "The rise in the disability rolls and the decline in unemployment," *Quarterly Journal of Economics*, Vol. 118, No. 1, February, pp. 157–205.

Barrow, Lisa, 2004, "Is the official unemployment rate misleading? A look at labor market statistics over the business cycle," *Economic Perspectives*, Federal Reserve Bank of Chicago, Vol. 28, No. 2, Second Quarter, pp. 21–35.

Barrow, Lisa, and Kristin F. Butcher, 2004, "Not working: Demographic changes, policy changes, and the distribution of weeks (not) worked," Federal Reserve Bank of Chicago, working paper, No. WP-2004-23.

Blinder, Alan S., 1973, "Wage discrimination: Reduced form and structural estimates," *Journal of Human Resources*, Vol. 8, No. 4, Autumn, pp. 436–455.

Carpenter, Christopher S., 2006, "The effects of employment protection for obese people," *Industrial Relations*, Vol. 45, No. 3, July, pp. 393–415.

Cawley, John, and Sheldon Danziger, 2004, "Obesity as a barrier to the transition from welfare to work," National Bureau of Economic Research, working paper, No. 10508, May.

Cutler, David, and Elizabeth Richardson, 1997, "Measuring the health of the U.S. population," *Brookings Papers on Economic Activity: Microeconomics*, Vol. 1997, pp. 217–282.

Juhn, Chinhui, Kevin M. Murphy, and Robert H. Topel, 2002, "Current unemployment, historically contemplated," *Brookings Papers on Economic Activity*, Vol. 2002, No. 1, pp. 79–135.

Lakdawalla, Darius, Jayanta Bhattacharya, and Dana P. Goldman, 2004, "Are the young becoming more disabled?," *Health Affairs*, Vol. 23, No. 1, January/February, pp. 168–176.

Oaxaca, Ronald, 1973, "Male–female wage differentials in urban labor markets," *International Economic Review*, Vol. 14, No. 3, October, pp. 693–709.

Philipson, Tomas J., and Richard A. Posner, 1999, "The long-run growth in obesity as a function of technological change," National Bureau of Economic Research, working paper, No. 7423, November.

Avoiding a meltdown: Managing the value of small change

François R. Velde

Introduction and summary

In 2007, the American bald eagle, a symbol of our nation, was removed from the threatened species list. But another American icon (or two) might well take its place on the list. On December 14, 2006, the United States Mint announced new regulations "to limit the exportation, melting, or treatment" of the American penny and nickel coins. The purpose of these regulations is "to safeguard against a potential shortage of these coins in circulation." The regulations make it illegal to export, melt, or treat one-cent and five-cent coins of the United States, except in some cases or with the Secretary of the Treasury's explicit permission.[1]

Our pennies and nickels, it turns out, are threatened with extinction by melting. Why that is the case, and what can be done about it, is the subject of this article. As inflation erodes the value of money, a coin of a given denomination (say, one cent or five cents) loses value. But coins are made of a physical material whose intrinsic value is usually low relative to the value of the coin, yet not negligible. Every now and then, we reach a point where the market value of the coin (its purchasing power) drops close to or below the intrinsic value of the materials used to make it. Our pennies and nickels have now reached that point.

This has two consequences. One is that the Mint is producing these coins at a loss. It now costs 1.67 cents to make a penny and 5.97 cents to make a nickel. The other is that it can be profitable to melt down the coins and recycle their metal content. The Mint's regulations were announced because we are close to the melt-down point for pennies and nickels.

The problem we are now facing is infrequent but, in many ways, a very old one. Seven hundred years ago (in the statute of 1299), the Parliament of England enacted that "no good money of silver, of the king's coin or other, nor any silver in plate or otherwise, should go forth or be carried out of the Realm or out of the King's power into foreign parts without especial leave from the king" (Ruding, 1817–19, Vol. 1, p. 385). This was the first of many such prohibitions—sometimes under penalty of death.

These prohibitions were passed at a time when money was different from ours, that is, when it was made of precious metals like gold and silver. Our money does not derive its value from its intrinsic content, which should be immaterial. In this article, I will first explain how a medieval problem can reappear in modern times. I will provide a quick overview of the history of American coinage, highlighting earlier instances of such problems, in particular the coin shortages of 1964 and 1965, and what solutions were adopted then. I will then discuss possible remedies to our current situation.

Historical background

The economy needs money to operate. Money is commonly described as having two functions: a unit of account in which prices and obligations are denominated and a medium of exchange in actual transactions. The two functions are logically distinct, but typically the unit of account has been tied to an actual medium of exchange. Coined money—that is, standardized quantities of metal shaped into a convenient form for everyday use—was introduced in Europe in the sixth century BC, and has served as a medium of exchange for almost all of subsequent history; and the unit of account has consequently been tied to the metal or metals coined.

In medieval Europe, where our modern system has its roots,[2] the metal was silver; the coin was the penny, made of silver alloyed with a little copper for

François Velde is a senior economist in the Economic Research Department at the Federal Reserve Bank of Chicago. *The author thanks his colleagues in the Economic Research Department for helpful comments.*

convenience. Governments set the standards by deciding how much metal went into a penny. The quantity of money was determined by the private sector, in the following way. If more money was needed, metal was brought to the mint and transformed into new coins, usually for a fee called seigniorage. If less money was needed, money was melted down and the metal turned to other uses. The signal for minting or melting was given by the price level (the inverse of the value of money): If silver in the form of coins was too cheap, it was profitable to turn it into bullion; if it was too expensive, it was profitable to sell bullion to the mint and acquire new coins. These two actions (and the equivalent actions of importing and exporting coins) served to regulate the price level.

The system worked well with one coin. But the growing needs of trade led to the introduction of larger silver coins and later even more valuable gold coins; and with multiple coins, the system does not work as well. The reason is that smaller coins are more expensive to make, in proportion to value, than larger ones. Mints were not subsidized and had to recover their production costs. This created a wider gap between minting and melting points for small coins than for large coins: The value of small coins had to go up higher before minting new ones became profitable (net of production costs). This led to a dilemma. If the mint bought silver for the same nominal price whether it paid in large or small coins, small coins had to contain less silver relative to their value and large coins might disappear and be melted down for their content. If the mint made all coins full-bodied, but charged more for small coins, large coins would be produced but not small coins, even when they were needed.

The Middle Ages were plagued with difficulties in maintaining an adequate supply of all denominations (Sargent and Velde, 2002). One common response to a shortage of one denomination was to prohibit the melting or exporting of the coins in short supply. The English statute of 1299 was an early example. It was followed within a few years by many other such statutes—a clear indication that such measures were difficult to enforce and had limited effect.

Another short-term solution was to debase the coin in short supply. Debasing a coin meant reducing its intrinsic content—for example, putting less silver in each penny. For a given market value of silver, debasement of one denomination can make it profitable to mint it again. In the case of medieval England, the cycle of melting prohibitions begun in 1299 led to a debasement of silver money in 1343. A debasement would restore the supply of the scarce denomination for a while, but inevitably shortages reappeared and further debasements followed. This repeated process led over time to pennies containing less silver and more copper, so much so that by the late eighteenth century, British (and American) pennies were made of pure copper.

A long-term solution was to return to the single-coin system, preserving the traditional minting and melting mechanism for one large gold coin and making the other coins token—that is, worth substantially more as money than as metal. The large coin pegged the value of the unit of account to a particular commodity, as in any commodity money system. Smaller denominations, however, were fiduciary; that is, their value in circulation was significantly higher than that of their intrinsic content. Their value came not from their content, but from a policy of convertibility: The authorities stood ready to exchange subsidiary coinage for gold coins, and vice versa. The provision of token coins was then left to the government, which bought and sold token coins on demand and made a profit from the substantial difference between face value and content. This is called the gold standard, and it became the norm, after much experimentation, in most countries by the end of the nineteenth century.

The U.S. monetary system, which Congress has sole power to regulate, began in 1792 as a bimetallic system.[3] This is a system in which silver and gold coins are provided by the minting and melting mechanism, and both coins play the same role as anchor of the monetary system. The founding fathers did not innovate at all in monetary matters. The bimetallic system was commonplace in Europe (though not in the mother country of Great Britain). The very mixed record of paper money during the colonial era, as well as the decidedly disastrous experience of the Continental money issued to finance the American Revolution, had predisposed the U.S. government to adhere to a commodity money system throughout the denomination structure. Even the smallest coins, the cent and half cent, were made of copper but were not token, which led to various problems. Copper was not that valuable and a cent's worth of copper was inconveniently large. Also, the world price of copper was volatile (because of its military uses), and it was difficult to maintain the cent at a fixed parity of 100:1 with the silver dollar. A similar problem arose from fluctuations in the relative price of gold to silver, which led to periods when no gold coin or no silver coin was minted and which prompted one debasement in 1834.

Prompted by the same forces as other countries, the U.S. gradually moved to a gold standard by making the smaller coins token. The first step was in 1853, when the silver content of quarters and dimes

was reduced relative to the silver content of the dollar coin. This proved difficult to enact, as there was reluctance on the part of many legislators to issue token money. Then, in 1873, silver dollars ceased to be minted on demand. The market value of silver fell substantially so that silver dollars became tokens too. The U.S. formally adopted the gold standard in 1900. Smaller coins were made of silver (the quarter and dime), nickel and copper (the nickel), or a copper alloy (the penny). The value of the silver in a quarter was around 10 cents. A quarter was worth 25 cents because the U.S. Department of the Treasury was always willing to exchange 40 of those coins for a gold $10 coin.

When the Federal Reserve System was created in 1914, the U.S. remained on a gold standard because Federal Reserve notes were redeemable on demand into gold at a fixed parity of $20.67 per ounce. The Great Depression, as well as the perceived need to increase the money supply without constraints to stimulate the economy, led to drastic changes. The gold content of the dollar was reduced by 40 percent, private holdings of gold by U.S. citizens were prohibited, and the Federal Reserve notes ceased to be redeemable on demand. The U.S. was on its way to a fiat money system (one in which money has value by fiat, that is, because the monetary authority or the government decrees it). After World War II, the Bretton Woods system restored a semblance of the gold standard, with foreign currencies convertible into dollars and dollars convertible into gold for foreigners. This lasted until 1971, when President Nixon closed the gold window and permanently severed the tie between the dollar and any commodity.

What about smaller denominations? In 1934, the Silver Purchase Act was passed, requiring the Treasury to purchase silver with the goal of reaching either a market price equal to its "monetary price" of $1.29 or a certain proportion of the monetary stock. The reasons for this action were complex: The issue of silver certificates in exchange for the silver purchased was to provide an additional avenue for increasing the money supply. Also, strong pressures from western states where silver was mined played a role in the legislation.

The market price of silver in late 1933 was 44 cents an ounce, and during the following years, the U.S. Department of the Treasury bought silver at above-market prices, between 50 cents and 77.5 cents an ounce, and after 1946 at 90.5 cents an ounce, accumulating a stockpile of 3,200 million ounces. By 1955, however, the world price of silver had risen to the Treasury's purchase price, and the Treasury began selling its silver, as it was authorized to do under existing legislation. Prices remained pegged at the Treasury's price of 90.5 cents an ounce, and the stockpile of silver that was not held to back silver certificates dwindled until November 1961, when President Kennedy stopped the sales. The price of silver started rising, and it reached the monetary price in September 1963.

At that price, the metallic content of dimes, quarters, half dollars, and dollars was exactly equal to their face value. Anyone needing silver for industrial uses could readily buy it on the commodities market as bullion or buy it from the banking system in the form of coins and melt them down.[4] As world supply and demand factors kept exerting upward pressure on prices, the U.S. monetary stockpile was drawn down, in various ways, either by redemption of silver certificates or else by the United States Mint working overtime to meet the "demand" for quarters and dimes. In early 1963, Treasury officials estimated that their silver supply would last 20 years. But the demand for subsidiary coinage proved unexpectedly strong, and the Mint's annual production quadrupled from 1963 to 1964. This was attributed initially to the growing use of vending machines, but it became clear that much of this demand was speculative: The public was buying the Treasury's stockpile at $1.29 an ounce in expectation of exhausting it and seeing the market price rise above the value they had paid. The Senate held hearings on the question in April and August of 1964 but came to no conclusion. The Treasury conducted its own studies and recommended in February 1965 that the silver content of subsidiary coinage be reduced or eliminated. In the end, following the recommendation of the Treasury studies, President Johnson proposed to Congress new legislation in June: It was swiftly voted into law and signed as the Coinage Act of 1965.[5]

The new law provided for the minting of the quarters and dimes made of copper and nickel (or cupronickel) that we know. The half dollar was replaced with a 40 percent silver core clad in copper and nickel.[6] The new quarters were issued in November 1965, by which time the reports of coin shortages had disappeared; dimes and half dollars followed in March 1966.

At the signing ceremony on July 23, 1965, President Johnson made curious remarks: "Some have asked whether our silver coins will disappear. The answer is very definitely no. … If anybody has any idea of hoarding our silver coins, let me say this. Treasury has a lot of silver on hand, and it can be, and it will be used to keep the price of silver in line with its value in our present silver coin. There will be no profit in holding them out of circulation for the value of their silver content."[7]

Indeed, the government's intention was not to replace silver dimes and quarters with cupronickel dimes and quarters, but only to reduce global demand for silver by removing the United States Mint from the ranks of the buyers. But keeping the existing stock of silver dimes and quarters in circulation was possible only if the price of silver did not rise above $1.29 an ounce. To achieve this, the Treasury had two means. One was its large stockpile of silver. The other was the authority given by the Coinage Act of 1965 to prohibit the melting and exportation of coins when necessary. The Treasury used both means in succession. First, for two years it sold silver at $1.29 an ounce, the price at which a quarter's content was worth 25 cents. The silver stockpile went from 1,200 million ounces in 1964 to 350 million in 1967. Then, using its new powers under the Coinage Act of 1965, it banned the melting, treatment, and export of silver dimes and quarters on May 20, 1967. Soon after, the Treasury stopped supplying silver at a fixed price on July 14, 1967, the day on which silver became "just another metal."

The prohibition met with some negative reactions in Congress, where two representatives introduced bills to repeal it, without success.[8] Although the ban was enforced and resulted in several indictments,[9] it did not prevent the disappearance of silver quarters and dimes from circulation. Silver half dollars had virtually disappeared from circulation by early 1966, and there were already reports of "culling" by consumers, that is, people picking out silver coins from their change and paying out only clad quarters.[10] By June 1968, the Treasury was itself melting silver quarters in its vaults, using new electronic sorting machines.[11] In August 1968, it was reported that dealers were paying 12 percent above face value for silver quarters and dimes. The combined forces of the Treasury and private speculators rapidly removed the silver coinage from circulation, making the ban moot. It was lifted in June 1969.

The provisions of the Coinage Act of 1965 were used a second time—this time to protect the penny. The peg to gold had ended in August 1971. Inflation was rampant, and commodity prices were exploding. On April 1, 1974, the price of copper reached a record of $1.40 per pound. At the time, 154 pennies contained one pound of copper. Although copper prices fell back somewhat, the demand for pennies rose to suspiciously high levels. The Treasury concluded that hoarding was under way in expectation that it would become profitable to melt pennies, and it announced the ban on April 18, 1974.

A few months later, Public Law 93-441 (31 USC 5112(c)) granted to the Secretary of the Treasury the power to change the proportion of zinc and copper in pennies to ensure adequate supplies. This gave the Treasury the option to replace copper with zinc in the composition of the penny, at its discretion. Copper prices stayed below the penny's melting point in subsequent years, so the ban was lifted in June 1978 without any further action. Soon, however, copper prices rose again and hit another record of $1.44 per pound on February 12, 1980. The Treasury briefly considered another ban, but instead used its statutory authority to change the composition of the penny, almost reversing the proportions. The Mint announced in June 1981 that, instead of 95 percent copper and 5 percent tin and zinc, pennies would be primarily zinc with a coating of copper; production started early the following year. As in 1965, no effort was made to retire the older coins: They were allowed to remain in circulation side by side with the new pennies.

Most people do not know that all pennies are not the same. Lincoln's profile has been unchanged since 1909. But take a penny dated 1983 or later and scratch its surface; you will see the shiny white zinc underneath the copper coating. As for the nickel, its size and composition have not changed since 1866. The effort to maintain the outward appearance of the coinage suggests the importance of habits in our attitudes toward coinage and currency.

The current situation

Between 1982 and 2004, the price of copper surged to the level of $1.50 per pound a few times, briefly. But in late 2004 it reached that level once more and has not come down since. Other commodities have surged in value as well, notably zinc and nickel. Table 1 shows the current value of the metal contained in U.S. coins.

The values shown in table 1 do not properly measure the profit to be made by melting down the coins. It would be necessary to subtract melting and refining costs (scrap copper is worth about 20 percent less than high-grade copper whose price is used in table 1). Collecting and shipping the coins for melting would impose additional costs, and those costs would be relatively larger for the smaller denominations, since digging a penny or a nickel out of a sofa requires the same effort.

Nevertheless, in 2006 some businesses became interested in the activity and inquired with the Mint about the legality of melting down coins. One firm in a midwestern state even began buying pennies from banks and sorting them to extract pre-1982 copper pennies. When the regulations were issued in December 2006, the Treasury had good reason to think that melting pennies and nickels was close to being profitable.

TABLE 1

Intrinsic value and composition of U.S. coins, 2007

Coin	Composition (percent of metal)	Intrinsic value (percent of face value)
Penny	95 zinc, 5 copper	69.7
Penny (pre-1982)	95 copper, 5 zinc and tin	209.5
Nickel	75 copper, 25 nickel	136.2
Dime, quarter, and Susan B. Anthony dollar	75 copper, 25 nickel	20.9
Golden dollar	88.5 copper, 2 nickel, and 3.5 manganese	5.7

Note: These are data as of November 14, 2007.
Sources: Author's calculations based on data from the United States Mint and Haver Analytics.

The nature of the problem

This brief historical overview frames the problem, which is something of a paradox.

A fiat money system is one in which money has value by fiat, that is, because someone said "let it be so." Economists like to describe money in their models as "intrinsically useless pieces of colored paper" because the challenge for monetary economics is to explain the value of such objects. For objects that are not intrinsically useless, we have standard price theory. For claims on objects that are not intrinsically useless, we have finance theory.

Since at least 1971, the U.S. has operated under a pure fiat money system, in which the intrinsic value of the objects used as a medium of exchange should not matter. This is in stark contrast with the commodity money regime of 1900. In that regime, the intrinsic content of coins provided a floor below which the value of coins could not fall, and minting on demand provided a ceiling above which it could not rise. The gap between floor and ceiling was usually fairly small. Under a fiat money regime, the ceiling is removed, as there is no minting on demand. The floor is normally of no consideration because no one pays much attention to the content of coins (copper pennies and zinc pennies circulate at par, although the content of the former is twice as valuable as the content of the latter). The stock of money, and its value, is determined not by minting and melting, but by the monetary authority's policy. In this respect there is no difference between notes and coins. The value of a dollar bill has nothing to do with its alternative uses as wallpaper or insulating material. Pennies and nickels are like notes, except they are made of something more durable than paper.

Now that all our currency is fiduciary (that is, with a market value higher than the intrinsic value), the market value of the tokens we use in physical transactions should be of no consequence to their value. The problem of small change was a difficult one to solve under a commodity money regime, but in a fiat money regime shortages of small change should not occur. The value of pennies and nickels has reached the floor set by their intrinsic content. We are printing our money on needlessly expensive material.

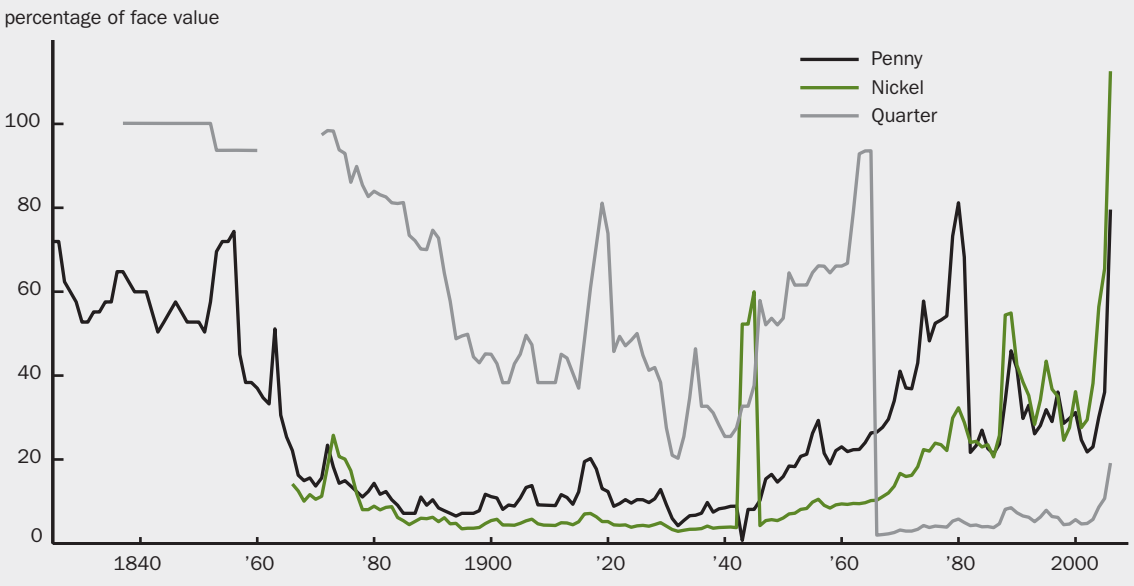
The historical overview also shows that this problem is not new. Figure 1 shows the value, as a percentage of face value, of the intrinsic content of coins minted every year since 1825 for three denominations. For all three types of coins (the penny, nickel, and quarter), there is over time a general upward trend; every time the value comes close to 100 percent, it becomes necessary to change the composition, which has the effect of abruptly lowering the value of the intrinsic content. The quarter began at 100 percent because it was a full-bodied coin, but in 1853 it became a subsidiary coin, overvalued relative to its silver content.[12] Figure 1 clearly shows what happened in 1965, when its composition was changed from silver to cupronickel. The line for the nickel displays a sharp uptick during World War II. At that time, nickel being needed for the war effort, nickels were made of silver. These coins swiftly disappeared in the early 1960s when the price of silver began to rise. Finally, the penny's line falls sharply in 1982 with the switch from copper to zinc.

The authority that the Secretary of the Treasury is using today to prohibit the melting and exportation of pennies and nickels was granted during the shortage of quarters and dimes in 1964–65. This authority was used to protect pennies in 1974. In each instance when the intrinsic value of the coin exceeded its face value, the long-term solution was to change the composition of the threatened coin.

Logic suggests, and history shows, that prohibitions on melting will not solve the problem. If it is really profitable to melt pennies or nickels, people will do it. The ban imposed in 1967 was lifted in 1969 because the coins it was designed to protect had disappeared. Such stopgap measures at best increase the costs of melting by a small amount—the probability of being caught times the penalties imposed. Devoting enough law enforcement resources to increase the probability of catching penny smelters hardly seems worthwhile. Alternatively, speculators can simply hoard the coins and incur time and storage costs as they wait for the regulations to be repealed. Those costs are real, but

FIGURE 1

Intrinsic value of U.S. coins, 1825–2006



Note: No quarters were produced from 1861 through 1870.
Sources: Author's calculations based on data from the United States Mint; and the U.S. Department of the Interior, U.S. Geological Survey.

they are modest compared with potential movements in commodity prices.

What drives this long-term trend in the intrinsic content of coins? Inflation is the answer. Although it was not much of a force in the nineteenth century (the price level was about the same in 1913 as in 1825), in the twentieth century it has been the main culprit. Money steadily loses its value relative to other goods, including the goods with which it is made. In other words, the floor on the value of coins is always creeping up, however slowly. In countries with high levels of inflation, the process can be rapid, and coins become obsolete in a matter of a few years. Our relatively low inflation in the U.S. means that these problems occur relatively infrequently, but they do occur.

The upward trend can be accelerated if metals rise in price faster than other goods. Figure 2 plots the real price of several metals that have been used in coins, deflated by the Consumer Price Index. The evidence is rather mixed. For some metals, such as aluminum, the secular trend is clearly downward. For other metals, there are long cycles—for example, the rise in the 1970s and the fall in the 1980s and 1990s. Since 2000, however, all metals have shown a sharp increase. The recent surge in commodity prices may arguably be speculative, and prices could well come down again, lowering the floor for a while. But as long as inflation is positive, the real value of a penny (which is always $0.01 in nominal terms) will fall relative to goods and services. When zinc replaced copper in the manufacture of pennies in 1982, the respite gained was relatively brief, since zinc was only half as costly as copper. Since zinc pennies were introduced, the real value of the penny (as measured by inflation) has fallen by half. Even if commodity prices stabilize, a 2 percent annual inflation rate will reduce the real value of the penny by another one-third over the next 20 years, and the problem will inevitably return unless another metal is found to replace those used in pennies and nickels.

Replacing the metal is not easy. As the law currently stands, the United States Mint has no authority to change the composition of the nickel and can only use copper and zinc for pennies. The Mint is nevertheless investigating alternatives. Finding a cheap metal is not enough: It must be easy to mint and must not present health risks, be allergenic, or wear out too quickly in circulation. Other countries, such as Canada, the United Kingdom, and those of the eurozone, have found steel a convenient substitute for other metals in the one-cent coin. Steel was used for the U.S. penny during World War II and was considered as an alternative in the 1970s. It has the advantage of being cheaper than other metals that have been used historically, such as aluminum (which was also considered in the 1970s), tin, and lead. New Zealand's

FIGURE 2

Real price of metals used in U.S. coins, 1825–2006

Notes: The data plotted are three-year moving averages. The values are deflated by the Consumer Price Index. There are no data available for silver from 1860 through 1871.
Sources: Author's calculations based on data from the U.S. Department of the Interior, U.S. Geological Survey; and Carter et al. (2006).

coinage now consists solely of steel cores, plated for aesthetic reasons with other metals and produced by the Royal Canadian Mint.

Even if a suitable metal is found, however, it will be difficult to produce pennies without taking a loss because production costs other than the metal were already 66 percent of face value in 2004 (see table 2). The Royal Canadian Mint is able to produce its penny for 0.8 cents.[13]

Should we eliminate the penny?

A simpler alternative is to let the penny melt out of existence. After all, do we need the penny?

The penny's role in our economy is not as a medium of exchange. There is nothing that a penny buys: Dime stores have long ago been replaced by dollar stores. Almost no coin-operated machinery accepts it.[14] We don't even use it truly to make change. It is merely a symbolic counter to simulate remainders of a division by five in retail transactions. When I buy a cup of coffee and the price comes out to $1.98, I give two dollar bills, the cashier takes two pennies from the saucer next to the register and hands them to me, and I return them to the saucer. The transaction is the same as if the cashier rounded to $2.00, except for a little side game between me and the cashier involving copper-colored tokens.

That I and the cashier are willing to give away the pennies in the saucer suggests that the penny isn't worth much. One way to see this is to measure the

TABLE 2

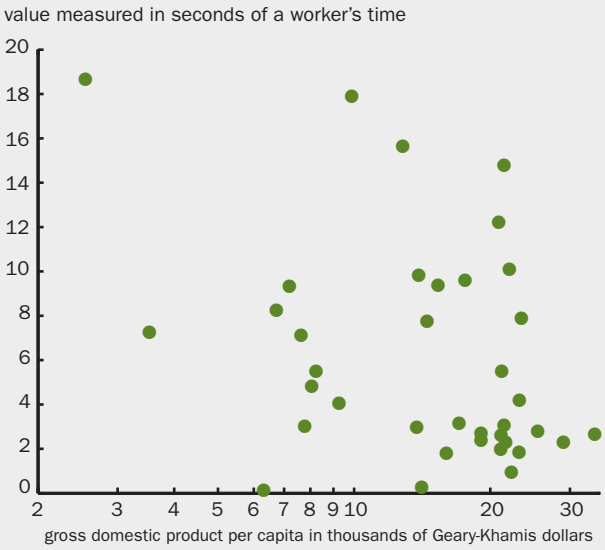
U.S. coin production costs and profits, 2004

Coin	Costs			Profits
	Metal	Other	Total	
	(- - - - percent of face value - - - -)			*($millions)*
Penny	27	66	93	2
Nickel	56	35	91	6
Dime	9	22	31	170
Quarter	9	20	29	424
Half dollar	9	25	34	2
Golden dollar	2	19	21	4

Notes: Metal cost is based on average metal prices for 2004. The profits are calculated from coin production numbers for 2004.
Sources: Author's calculations based on data from the United States Mint, *United States Mint Annual Report 2004*; and the U.S. Department of the Interior, U.S. Geological Survey.

FIGURE 3

Value of the smallest circulating coin compared across countries, 2003



Notes: The sample includes the OECD (Organization for Economic Cooperation and Development) countries plus other European countries. For details on Geary-Khamis dollars, see Maddison (1995).
Sources: Author's calculations based on data from the Organization for Economic Cooperation and Development, International Labor Organization, national mint websites, and Maddison (1995).

penny with the value of time. Median weekly earnings for wage earners and salaried workers are $675. Assuming a 40-hour workweek, it takes most U.S. workers no more than two seconds to earn a penny. Rounding transaction prices to the nearest five cents would save more than the time we spend fishing for pennies in our pockets or wallets.

A comparison with other countries is instructive. Figure 3 compares the values of the smallest circulating coins in about 30 countries—mostly the OECD (Organization for Economic Cooperation and Development) countries plus other European countries. The values of each coin are again measured in the time it takes to earn it at the average wage in manufacturing. The values are plotted as a function of gross domestic product (GDP) per capita, measured in Geary-Khamis dollars (Maddison, 1995). There seems to be a small negative relationship. However, this relation is not very robust and is largely due to the recent adoption of the euro as the common currency by the relatively rich European countries (the same figure in 1999, right before the introduction of the euro, shows no significant relationship between the value of small coins and GDP per capita).

What figure 3 does show is that there is a wide range across countries in terms of the value of their smallest denomination. That is in part because, in recent years, a number of countries have abandoned their smallest denominations. In Australia and New Zealand, whose dollars are comparable in value to the U.S. dollar, one-cent coins were also made essentially of copper. In 1987, the rise of copper prices made the one-cent coin unprofitable to mint. Instead of changing the content, New Zealand stopped producing its one-cent and two-cent coins (worth about 0.5 cents and one cent in U.S. currency, respectively) in March 1989, and they ceased to be legal tender in April 1990. The coins were bought back by the Reserve Bank of New Zealand and melted down for scrap metal.[15] Australia followed suit, stopping production of the coins in August 1990 and issuance in February 1992.[16] New Zealand went further in 2006: Existing five-cent, ten-cent, 20-cent, and 50-cent coins ceased to be legal tender, and all but the five-cent denomination were replaced with smaller and cheaper coins of plated steel.

In the eurozone, the smallest euro denominations are the one-cent (currently worth about 1.5 cents in U.S. currency) and two-cent coins. Each country can mint its own coins (with a common European obverse and nationally designed reverse), and all coins are legal tender throughout the eurozone. Two countries, the Netherlands and Finland, opted not to issue one-cent and two-cent coins at all, and they officially encourage rounding to the nearest five cents within their borders. Outside of the eurozone, the Czech Republic and Slovakia have recently eliminated their two smallest coins, and Hungary plans to do so next year.

The penny is disappearing of its own accord in economic terms. Various interest groups (for example, zinc producers, charities, and the state of Illinois) can point to continued support for the penny shown in polls.[17] But the United States Mint's annual output of pennies, nickels, dimes, and quarters as a ratio of GDP tells a different story (see figure 4). While the relative importance of 25-cent coin output has been stable over the past 30 years, that of the other coins has been declining steadily. Relative to GDP, the output of pennies is 12 percent of what it was in 1982. The trend is not much better for the nickel.

So a penny isn't worth much and the quantities produced are declining relative to GDP, but we still

FIGURE 4

Ratio of U.S. coin production to gross domestic product, 1946–2006

Note: The data plotted are three-year moving averages.
Sources: Author's calculations based on data from the United States Mint and U.S. Bureau of Economic Analysis.

produce a lot of them. Since 1982, the Mint has produced 910 pennies for every man, woman, and child in America. It estimates that 100 billion pennies currently circulate. In 2006, the Mint used 20,000 tons of zinc, worth $60 million, to produce pennies. Even if the Mint (and the taxpayer) were not losing money on this activity, it would be fair to ask whether all that zinc might be put to better use than manufacturing throwaway tokens.

The declining value of the penny is not a temporary phenomenon. It is a trend driven by several factors. One, noted previously, is inflation. The penny has been part of our denomination structure since the beginning, in 1792, but the price level has gone up by a factor of 20 in the past century: A penny today is worth one-twentieth of a penny before World War I. If people got by without coins as small as 0.05 cents back then, we can probably do so today. A second factor is that, even in the absence of inflation, a penny means less over time because we are becoming richer. As productivity grows, a penny will be worth ever less of our time because our time is more productive. A third factor is the replacement of cash (coins and notes) by other means of payment, notably electronic ones. Just as there was a boom in the demand for coins in the 1950s and 1960s because of the spread of coin-operated machinery, we can expect technological change to affect the demand for coins in the future.

These factors together tell us that the penny will disappear sooner or later, as did the farthing (one-quarter of a penny) and the ha'penny (one-half of a penny) of medieval England, and our own half cent, last minted in 1857.

Moreover, the experience of other countries suggests that there are few problems involved in doing so. The Reserve Bank of New Zealand has not found any evidence of inflation or upward rounding since it withdrew its one-cent and two-cent coins. The Royal Canadian Mint recently published survey results indicating that small retailers were vastly in favor of removing the penny, and consumers were split on the issue.

Current legislative proposals

As I noted earlier, the solution to our problem of small change is constitutionally vested in the hands of Congress, and some legislation is on the agenda.

Two bills were introduced in Congress in early August 2007.[18] Both bills confer on the Secretary of the Treasury the power to "prescribe the weight and the composition" of existing denominations, considering "such factors that the Secretary considers, in the Secretary's sole discretion, to be appropriate." A third

bill introduced in October 2007 includes a similar provision.[19]

Delegating such power to the Secretary of the Treasury would represent a significant change. Ever since the Coinage Act of 1792,[20] Congress has retained for itself the exercise of its constitutional powers to "coin money, regulate the value thereof." There were good reasons for the founding fathers to assign such powers to Congress. Under a commodity money system (the only system they could conceive for our country), setting the weight and composition of coins is the essence of monetary policy and is therefore an extremely important power. Recent European history, with which they were familiar, gave them reason to be wary of handing over monetary policy to the executive branch.

But things have changed. The composition of coins is not central to monetary policy anymore. Under a fiat system, it is a purely technical issue, whose only potential consequence for the legislature is the profit or loss made on coining.

Profit on coinage, of course, is not negligible. Table 2 (p. 23) shows that, on some coins, the profits can be substantial. The high figure for the quarter reflects the success of the "state quarters" program, which has generated $3.2 billion in "above-average" profits on this denomination over eight years.

But profits can rapidly turn into losses. The United States Mint made a small profit ($5 million) on pennies and nickels in fiscal year 2005, but this turned into a loss of $33 million in fiscal year 2006 and a loss of almost $100 million in fiscal year 2007.

Congress therefore retains an interest in the issue of coin composition, but it could nevertheless delegate the details to the executive branch (namely, the U.S. Department of the Treasury) because the issue is purely technical and because action in the executive branch will be timelier than passing new legislation each time.[21]

A medieval solution to a medieval problem

In a recent *Chicago Fed Letter*, I made a different proposal.[22] Starting from the observation that there are many pennies in circulation but they are not really needed as one-cent coins and inspired by medieval debasements, I proposed that the prohibition on melting should be repealed and that pennies should henceforth be worth five cents.

In this proposal, the existing nickels would disappear and be melted down, which seems likely to be their fate under any conceivable proposal. Pennies would then be recycled as five-cent coins, avoiding the need to design and produce a new coin (a lengthy process). Since the Mint has produced about seven times as many pennies as nickels in the last 20 years, there should be enough pennies to serve as five-cent coins for a while.

The new value would be easily established by the monetary authority standing ready to exchange 20 pennies for a dollar bill, instead of 100 pennies presently. It is true that vending machines and other coin-operated equipment currently accepting nickels would have to be modified to accept pennies as five cents. But such modifications may be unavoidable if the nickel in its current form is doomed.

I call this a medieval solution because medieval debasements were sometimes carried out in this manner. When a coin was threatened by melting, as is now the case with our penny, there were two ways to debase it: One was to mint it with less metal than before, and the other was to increase its face value. Thus, in 1269 Venice increased the face value of its grosso coin from 26 to 28, and again in 1282 to 32, each time leaving its composition unchanged. As I recently found out, the idea also has precedent in U.S. history. During the 1965 silver coinage crisis, Congressman Craig Hosmer (a Republican from California) proposed to "arbitrarily double the value of existing silver coins" in order to save them from being melted down.[23]

The proposal would require everyone to ignore the inscription on the penny that says "one cent." But there is also precedent for U.S. coins being worth more than what is written on them. In 1834, when the gold–silver ratio was adjusted, half eagles minted before that date and bearing the inscription "5 D" (five dollars) were declared to be "receivable in all payments at the rate of 94 and 8/10ths of a cent per pennyweight," which works out to $5.33 for a full-weight coin.[24] This was nothing else than a debasement, albeit a relatively modest one.

Would such a measure be inflationary? The estimated stock of pennies is 100 billion, so increasing their value to five cents would add $4 billion to the money supply, which represents 0.5 percent of the monetary base or 0.3 percent of M1 (a monetary aggregate composed of currency and demand deposits). This is a modest addition. The average monthly increase in the monetary base over the past three years has been about $2 billion; the monthly standard deviation of M1 is about $6 billion over the same period. Thus, an addition of $4 billion would fall well within the range of typical monthly variations in the money supply. The one-time increase would also be offset by reduced issues of other coins and thus unlikely to have a noticeable impact.

Conclusion

To prevent a shortage of small change, the U.S. Department of the Treasury recently enacted regulations to prohibit melting and exportation of pennies and other coins. The threat of shortage arises because pennies and nickels are made of inappropriately expensive material, and there is or soon will be a profit to be made from transferring their content to alternative uses.

There is $1 billion worth of resources sitting in cash registers, jars, and sofas across the United States. It makes little sense to keep replenishing them, and the regulations hold little promise of forestalling the inevitable very long. The traditional solution since medieval times is to "debase" the threatened coin, that is, make it of a cheaper material or assign it a higher face value, either of which requires congressional action. But the current situation may well prompt a more general debate on whether such small denominations are worth saving—a debate that is ongoing in many other industrialized countries.

NOTES

[1]The regulations became permanent on April 16, 2007, and now constitute 31 CFR Part 82 (*Federal Register*, April 16, 2007). By law (31 USC 5111 (d1)), the Secretary of the Treasury "may prohibit or limit the exportation, melting, or treatment of United States coins when the Secretary decides the prohibition or limitation is necessary to protect the coinage of the United States." One of the exceptions to the regulations allows Federal Reserve Banks and depository institutions to continue exporting coins for circulation in "dollarized" countries, such as Ecuador and Panama.

[2]The word "penny" itself goes back at least to the ninth century.

[3]See Carothers (1930).

[4]This neglects refining costs: Coins consisted of silver at 90 percent purity mixed with copper.

[5]The Coinage Act of 1965 is also known as Public Law 89-81 (79 Stat. 254).

[6]The silver core was abandoned in 1971; at the same time the Eisenhower dollar, also made of copper and nickel, was introduced to replace the silver dollar discontinued in 1964.

[7]Times Mirror Company (1965).

[8]Cabeen (1967).

[9]Three Manhattan jewelry technicians were arrested in December 1967. Three men were arrested near Tucson, AZ, with two tons of dimes and quarters and a small smelter in April 1968; two men were arrested in Brooklyn and arraigned in December 1968 (Laurence 1968; Dow Jones and Company, 1968a, b).

[10]Janssen (1966).

[11]Times Mirror Company (1968).

[12]The break in the quarter series in figure 1 is related to another shortage of small change—this one prompted by the introduction of fiat money in the form of "greenbacks," notes that were not redeemable into gold or silver. During the subsequent period of inflation, from 1861 through 1870, the dollar price of silver made it unprofitable to mint silver quarters, while existing quarters were hoarded or melted.

[13]Branswell (2007).

[14]One notable exception is the acceptance of pennies in automatic toll lanes on Illinois roads. Until a few years ago parking meters in downtown Hilo, HI, accepted pennies, but parking is now free.

[15]APN News and Media (1990).

[16]Glover (1992).

[17]Hagenbaugh (2006).

[18]HR 3330 and S 1986.

[19]HR 3956.

[20]1 Statutes at Large 246.

[21]The Reserve Bank of New Zealand is vested with the power to "determine the denominations, form, design, content, weight, and composition of its bank notes and coins," according to the Reserve Bank of New Zealand Act 1989, s. 25(2). Thus, the recent decision to abolish the five-cent denomination was taken by the Reserve Bank of New Zealand, without any legislation.

[22]Velde (2007).

[23]Foley (1965).

[24]4 Statutes at Large 699, section 3.

REFERENCES

APN News and Media, 1990, "Small coins worth $3m are sold as scrap," *New Zealand Herald*, April 26, p. 9.

Branswell, Jack, 2007, "Is penny lane a dead end?," *Montreal Gazette*, October 9, p. B1.

Cabeen, Richard McP., 1967, "Rules on handling silver coins may be due for a change," *Chicago Tribune*, August 27, p. F8.

Carothers, Neil, 1930, *Fractional Money: A History of the Small Coins and Fractional Paper Currency of the United States*, New York: John Wiley and Sons.

Carter, Susan B., Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright (eds.), 2006, *Historical Statistics of the United States*, Millennium ed., 5 vols., New York: Cambridge University Press.

Dow Jones and Company, 1968a, "Two men are seized, charged with illegal melting of coins," *Wall Street Journal*, December 5, p. 29.

Dow Jones and Company, 1968b, "Three men charged with melting coins to reclaim silver," *Wall Street Journal*, April 30, p. 9.

Foley, Thomas J., 1965, "Silverless and lighter coins asked by Johnson," *Los Angeles Times*, June 4, p. 1.

Glover, Richard, 1992, "To coin a phrase, coppers are a rum deal now," *Sydney Morning Herald*, January 30, p. 3.

Hagenbaugh, Barbara, 2006, "A penny saved could become a penny spurned," *USA Today*, July 7, p. B1.

Janssen, Richard F., 1966, "Gresham's law faced by Mint," *Wall Street Journal*, February 16, p. 1.

Laurence, Michael, 1968, "Better than money. Better than money?," *New York Times,* August 25, p. SM4.

Maddison, Angus, 1995, *Monitoring the World Economy, 1820–1992*, Paris: Development Center of the Organization for Economic Cooperation and Development.

Ruding, Rogers, 1817–19, *Annals of the Coinage of Great Britain and its Dependencies*, 3 vols., London: Nichols, Son, and Bentley.

Sargent, Thomas J., and François R. Velde, 2002, *The Big Problem of Small Change*, Princeton, NJ: Princeton University Press.

Times Mirror Company, 1968, "Silver quarters go to meet their melter," *Los Angeles Times*, June 27, p. E24.

Times Mirror Company, 1965, "New 'sandwich' coins bill signed by Johnson," *Los Angeles Times,* July 24, p. 3.

Velde, François, 2007, "What's a penny (or a nickel) really worth?," *Chicago Fed Letter*, Federal Reserve Bank of Chicago, No. 235a, February.

Corruption and innovation

Marcelo Veracierto

Introduction and summary

In this article, I illustrate how corruption can lower the rate of product innovation in an industry. This is important because, if many industries are subject to corrupt practices, the lower rate of innovation would result in a lower growth rate for the whole economy.[1] Actually, the view that corruption is closely related to economic development is widely held in practice: Poor African countries, such as Kenya and Zaire, are commonly believed to lose a considerable fraction of their gross domestic product (GDP) to corruption activities. Figure 1 illustrates the extent of this perception. It plots 2004 GDP per capita levels from the Penn World Table against the 2004 Corruption Perception Index constructed by Transparency International.[2] Since a Corruption Perception Index number close to zero indicates no corruption, figure 1 shows a clear negative relation between corruption and economic development.

While a negative correlation between corruption and GDP per capita levels is highly suggestive of an actual link, it is not conclusive evidence. It may be the case that corruption is closely related to other variables, such as political instability, the extent of violence, or the combativeness of unions, among other factors, and that these other variables are the ones generating poor economic development outcomes. In addition, GDP per capita levels may be affecting corruption levels and not the other way round. To complicate matters further, the negative correlation between corruption indexes and GDP per capita levels could be a mere artifact: It may well be the case that low GDP per capita levels are biasing the subjective perception of corruption reported by survey respondents. To disentangle the effects of corruption on economic development, further analysis is needed.

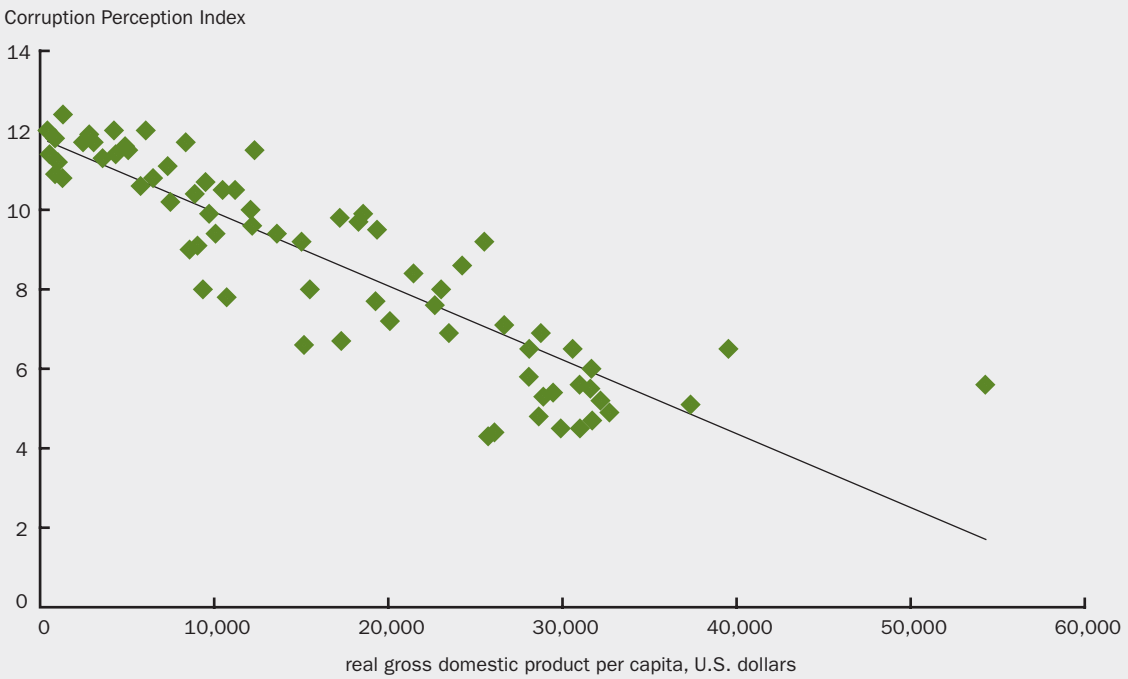
In this article, I provide theoretical grounds for pursuing such an analysis: In particular, I explore the strategic interactions between producers and corrupt officials. The basic corruption scenario considered involves three agents: an innovator, an incumbent producer, and a corrupt government official. The innovator wants to enter business by potentially paying a bribe; the incumbent producer wants to preclude the entry of the innovator by potentially paying a bribe; and the corrupt official decides on allowing the entry of the innovator based on the bribes received. Key elements of the game are that the government official can make successive take-it-or-leave-it bribe offers to the producers and that the central government can never verify the actual payment of a bribe (with some probability, the central government can detect that the entry permit was misallocated but cannot prove the actual amount of the bribe paid). Under these assumptions and within certain ranges, I show that the amount of bribes that the government official can collect can be very responsive to small changes in the probability of detection or in the penalties imposed. In fact, the bribe payments are shown to be a discontinuous function of those variables. Since the resources devoted to innovation are continuously and inversely related to the bribes that producers must pay, this means that the amount of resources devoted to innovation is a discontinuous function of the probability of detecting corruption and of the penalties imposed.

The rest of this article is organized as follows. In the next section, I discuss the related literature. Then, I describe the corruption game and characterize its solution. Next, I analyze the implications of the corruption game for innovation decisions. Finally, I draw some conclusions about my findings.

Marcelo Veracierto is a senior economist in the Economic Research Department at the Federal Reserve Bank of Chicago. *The author thanks Marco Bassetto, Craig Furfine, and seminar participants at the Federal Reserve Bank of Chicago for their comments.*

FIGURE 1

Corruption and real gross domestic product per capita



Notes: The Corruption Perception Index reported here is actually defined as 14 minus the Corruption Perception Index constructed by Transparency International. The transformation is made to associate low values for the index with low levels of corruption.

Sources: Author's calculations based on data from the University of Pennsylvania, Center for International Comparisons, Penn World Table; and Transparency International, Corruption Perception Index.

Related literature

Systematic empirical evidence about the relationship between corruption and economic development is hard to come by. A notable exception is the study by Mauro (1995). Using Business International Corporation's indexes on corruption, red tape, and efficiency of the judicial system over the period 1980–83 (now incorporated into the Economist Intelligence Unit), Mauro was able to estimate the direct effects of corruption on economic development. He found that corruption lowers investment, even controlling for other determinants of investment and endogeneity effects. The magnitude of the effect is quite significant. Mauro found that a one standard deviation improvement in the corruption index is associated with an increase in investment of 2.9 percent of GDP. This means, for example, that "if Bangladesh were to improve the integrity and efficiency of its bureaucracy to the level of that of Uruguay, its investment rate would rise by almost five percentage points, and its yearly GDP growth rate would rise by over half a percentage point" (Mauro, 1995, p. 705).

On the theoretical side, the literature has proceeded along two lines. One, following Becker and Stigler (1974), used a principal–agent approach. In particular, it focused on the incentives that the central government (the principal) can give a government official (the agent) to make him behave honestly. Another strand, following Shleifer and Vishny (1993), took the corrupt behavior of government officials as a given and analyzed the consequences that their behavior has on resource allocation. In this approach, corrupted officials are modeled as monopolistic suppliers of a government good (such as a passport, an import license, the right to use a road, etc.) that is supposed to be supplied at a prespecified price. The corrupt official overcharges the government good to maximize his total revenues.

More recently, Acemoglu and Verdier (2000) took a broader approach. They considered a static economy in which producers can choose to pay a cost in order to produce with a clean technology (otherwise, their production process pollutes the environment). The government wants to tax polluters and subsidize clean producers in order to reduce the associated negative externality. However, it must rely on officials to

inspect the producers and determine their pollution status. The officials are assumed to be corrupt: Through bribes they are able to grab an exogenous share of the surplus, which is assumed to be equal to the sum of the tax and the subsidy that the official can potentially charge. As a consequence, the government faces an important trade-off between taxation and corruption: It wants to tax polluters, but in order to detect them it must rely on corrupted officials that consume resources. In this environment, Acemoglu and Verdier (2000) characterize the optimal amount of taxation/corruption.

While Acemoglu and Verdier (2000) were able to analyze the optimal taxation/corruption policy of the government, in order to do so they had to simplify the interaction between the government officials and the producers to a reduced form. My contribution to this literature is to spell out that interaction in an explicit game and analyze its implications in detail. Since Djankov et al. (2002) report that there are large differences across countries in the regulation of entry and that this type of regulation is associated with sharply higher levels of corruption, I formulate the corruption game in the context of entry decisions to an industry.[3]

The corruption game

The corruption game is as follows. Consider the case of a product line that is supplied by a single producer—the incumbent. The value of supplying the product line is given by V. In addition, there is a potential producer that has just created a new product generation—the innovator. If the innovator is allowed to supply the new product, the incumbent will be driven out of the market. As a consequence, the innovator would obtain the value V and the incumbent would lose it. Entry is regulated: The innovator must receive permission from the government to enter business. The reason for the regulation is that the innovator may produce with a technology that pollutes the environment. The government is willing to grant the entry permit to the innovator only if the new production technology is clean. However, the government must send a government official to determine whether the new technology pollutes or not. Once the government official inspects the new technology, its pollution status becomes fully known to him. After the official learns the pollution status of the new technology, he must report it to the central government. If the official reports that the new technology pollutes the environment, the innovator is precluded from producing but faces no additional penalties.[4]

The government official is corrupt. He has the ability of misrepresenting to the government the true pollution status of the new technology. This allows him to try to extract a bribe, either from the incumbent or the innovator, in determining which report to make to the central government. For simplicity, I assume that the pollution status of the new technology is fully known to both the innovator and the incumbent. This means that once the government official inspects the new technology, its pollution status becomes common knowledge to the three parties—the incumbent, the innovator, and the official.

The government never observes the actual bribe payment received by the official. However, once the official makes his report, with probability φ the government independently learns about the true pollution status of the new technology. If the official is found to have granted an entry permit to a polluter, there are penalties involved. In particular, the official is fined pV, while the innovator is fined mV. If the official is found to have rejected an entry permit to a clean innovator, there are also penalties involved: The official is fined pV, and the incumbent is fined mV.

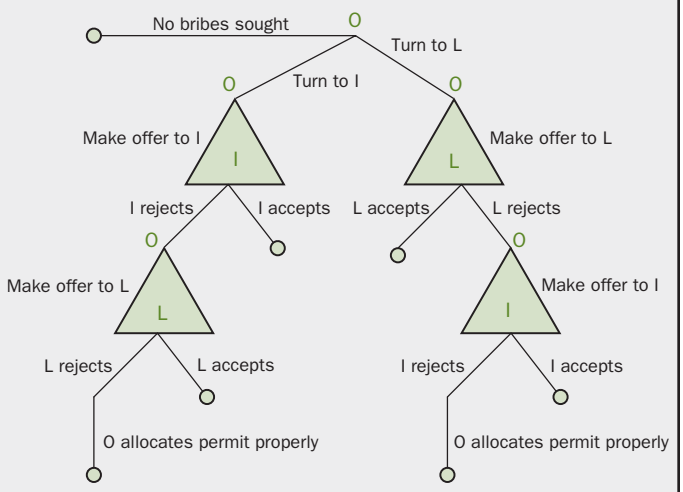
The official is assumed to be able to make take-it-or-leave-it offers. The key decision for the official is whether to request a bribe and from whom. In what follows, we will see that the best strategy for the official is to turn to the producer with the largest joint surplus and make him a take-it-or-leave-it offer. However, for the producer with the largest joint surplus to accept this bribe proposal, it must be credible that the producer with the second largest joint surplus would be willing to accept a bribe proposal if offered one. This will require the second largest joint surplus to be positive.

In principle, two possible scenarios must be considered: the scenario in which the innovator does not pollute the environment and the scenario in which the innovator does pollute the environment. However, the two scenarios are completely symmetrical. In each scenario there is a "legal" producer and an "illegal" producer. In the case that the innovator pollutes, the legal producer is the incumbent; in the case that the innovator does not pollute, the legal producer is the innovator. Moreover, the payoffs to each player in the corruption game only depend on whether the bribes are being extracted from the legal producer or the illegal producer (that is, the payoffs are independent of the actual identity of the producers). Given this symmetry, in what follows I consider a single corruption game that differentiates producers only according to their legal status, with the understanding that the identities of the legal and illegal producers are determined by the actual scenario taking place.

Figure 2 describes the sequence of moves for the corruption game. In the first stage, the government

FIGURE 2

Sequence of moves for corruption game



Notes: I refers to illegal producer; L to legal producer; and O to government official. See the text for further details.

official must decide between three alternatives: 1) not to seek bribes, 2) to initially seek a bribe from the illegal producer *I*, and 3) to initially seek a bribe from the legal producer *L*. In the case that the official seeks a bribe, he must decide how much to demand from the producer he initially turns to (the continuum of values for the bribe are represented as the base of the triangles in figure 2). If the bribe request is accepted, the game ends. Otherwise, the official turns to the second producer and decides how much to demand from him. The game ends after this point. If this bribe request is rejected, the official assigns the production permit to the legal producer, since he has nothing to gain otherwise. In what follows, I analyze the way that the corruption game is played.

First, observe that the government official always has a larger joint surplus to share with the legal producer than with the illegal producer. The reason is that the value of being the product leader is the same for both types of producers, but there are penalties involved if a deal with the illegal producer is subsequently detected. In addition, the value of not being the product leader is the same for both types of producers (in particular, it is equal to zero). This means that the government official will always want to extract bribes from the legal producer. However, for the legal producer to be willing to pay such a bribe, it should be credible that the government official would want to reach a deal with the illegal producer in a second round of negotiation. If this is not the case, the legal producer will reject any bribe request, since he knows that the government official will subsequently take the legal course of action.

Observe that the payoff to the illegal producer of reaching a deal with the government official is:

$$P_I = V - B_I - \varphi\,[V + mV],$$

where B_I are the bribes paid. This payoff is equal to the value of being the product leader net of the bribe payment minus the losses if the deal is detected, an event that happens with probability φ. Since the payoff to the illegal producer of rejecting the bribe is zero, the largest bribe that the government official would be able to extract from the illegal producer in a take-it-or-leave-it offer is given by:[5]

$$1) \qquad B_I = (1-\varphi)V - \varphi mV.$$

The payoff to the government official in this case is

$$P_O = B_I - \varphi pV = (1-\varphi)V - \varphi mV - \varphi pV.$$

That is, it is the maximum bribe that the government official could extract from the illegal producer minus the penalty pV times the probability φ of being caught by the central government.

The condition that this payoff P_O is positive reduces to

$$2) \qquad \frac{1-\varphi}{\varphi} > m + p.$$

If this condition is not satisfied, it would be in the best interest of the government official not to seek a bribe from the illegal producer. Hence, the legal producer would reject any take-it-or-leave-it offer made by the official and the legal course of action would be taken. If the condition in equation 2 is satisfied, the government official would be able to extract bribes from the legal producer, since it becomes fully credible that he would subsequently want to reach a deal with the illegal producer.

Observe that the payoff to the legal producer of reaching a deal with the government official is:

$$P_L = V - B_L.$$

That is, it is equal to the value of being the product leader net of the bribes paid. This payoff is nonrandom because if the central government independently learned about the pollution status of the innovator, it would conclude that the entry permit was correctly allocated (recall that the central government can never prove that a bribe payment took place). Also, the payoff to the legal producer of rejecting a bribe offer from the government official is φV, since with probability φ the illegal action will be detected by the central government and the legal producer will become the product leader. Hence, the largest bribe that the government official will be able to extract from the legal producer is:[6]

3) $B_L = (1 - \varphi)V.$

To summarize, the equilibrium of the corruption game is as follows. If the condition in equation 2 is violated, no bribes are paid. If the condition in equation 2 is satisfied, the government official extracts from the legal producer the bribes given by equation 3. In both cases, the official takes the legal course of action. Figure 3 provides an illustration of the equilibrium outcome.

Innovation decisions

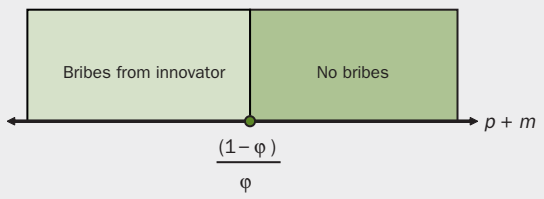
In this section, I describe in detail the industry in which the incumbent and innovator of the previous section operate. The purpose is to determine how corruption affects the industry's innovation rate.

The industry produces a product that comes in many possible qualities. At each point in time, there is a frontier version that dominates all previous ones. A single producer has the patent to this version. He drives all other producers out of the market and enjoys a profit flow equal to Π. However, he loses his leading position whenever an innovator enters business with a quality improvement. In this case, the incumbent is driven out of the market, and the innovator becomes the new industry leader, which provides him the profit flow Π.

Product innovations take place at an endogenously determined rate η. At every point in time there are a large number of potential producers (innovators) that invest in research and development (R&D) in order to create a new product generation. They all face a same cost function $r(\eta)$, which describes the costs of generating an arrival rate equal to η.[7] If an innovator succeeds in creating the new product generation, he can apply for an entry permit. If the entry permit is awarded, the innovator becomes the new industry leader. However, entry is regulated as in the previous

FIGURE 3

Equilibrium outcomes in parameter space



Note: See the text for further details.

section. In particular, a government official is sent to inspect the pollution status of the new technology. As a result, the official, the incumbent producer, and the innovator end up playing the corruption game described before. The probability that an entry application is inspected by a government official is equal to γ, while the probability that an innovation pollutes is equal to ξ.

The optimization problem of an innovator is then the following:

4) $\max \{\eta [\xi N_P + (1 - \xi) N_C] - r(\eta)\},$

where N_P is the value of being an innovator that pollutes and N_C is the value of being an innovator that produces with a clean technology. That is, the innovator chooses the arrival rate η to maximize the expected value net of R&D costs. The optimal innovation rate η is characterized by the following condition:

5) $r'(\eta) = \xi N_P + (1 - \xi) N_C .$

That is, the innovator equates marginal revenue to marginal cost. In what follows, I sketch the main properties of the optimal R&D investment decisions both from an individual point of view and at the industry level. The appendix provides a more detailed analysis.

To start with, observe that the marginal cost function r' is strictly increasing. Thus, given fixed values for N_P and N_C, there is a unique value of η that satisfies equation 5. While an individual innovator takes the values of N_P and N_C as given (since he is competitive), these values actually depend on the industry-wide innovation rate η^*. Moreover, they are strictly decreasing in the industry-wide innovation rate η^*. The reason is that given all other parameter values, an increase in η^* decreases the expected length of time over which a producer can retain the leadership of a product line (that is, it increases the

rate at which future innovators will drive him out of the market). Thus, the expected value

$$\bar{N} = \xi N_P + (1-\xi) N_C$$

in the right-hand side of equation 5 is strictly decreasing in η^*. At equilibrium, the industry-wide innovation rate η^* that innovators take as given (and that determines the expected value $\bar{N}$) must be identical to the one they choose from their individual perspective. That is, at equilibrium we must have that the innovation rate satisfies:

$$6) \quad r'(\eta^*) = \bar{N}(\eta^*).$$

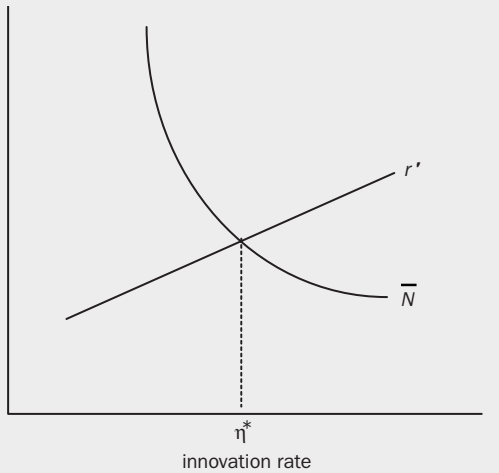
Since the left-hand side of equation 6 is strictly increasing in η^* and the right-hand side of equation 6 is strictly decreasing in η^*, there is a unique value of η^* that satisfies this equation. That is, there is a unique industry equilibrium. Figure 4 illustrates this equilibrium.

We are interested in how the equilibrium innovation rate η^* is affected by changes in different parameter values. While the appendix provides a formal analysis, the results are quite intuitive. We saw in the previous section that the penalties to the government official and illegal producer (p and m, respectively) affect whether bribes are paid or not but do not affect the magnitude of the bribes. In particular, if the condition in equation 2 is satisfied, bribes are paid. However, p and m do not enter equation 3, which describes the equilibrium bribes B_L that the government officials are able to extract from the legal producers. This means that as long as $p + m > (1 - \varphi)/\varphi$, the expected value $\bar{N}$ is independent of those penalties; but as soon as $p + m$ becomes equal to $(1 - \varphi)/\varphi$, the expected value $\bar{N}$ plummets because now producers become subject to bribes. Further, decreases in $p + m$ have no additional effects in $\bar{N}$. The implications for the equilibrium innovation rate are shown in figure 5. The curve $\bar{N}_1$ describes the expected value of innovating in the case in which there are no bribes (that is, when $p + m > (1- \varphi)/\varphi$), while the curve $\bar{N}_2$ describes the expected value of innovation when producers pay bribes (that is, when $p + m < (1 - \varphi)/\varphi$). Since $\bar{N}_2$ is lower than $\bar{N}_1$ for every value of η, it follows that the equilibrium innovation rate with bribes η_2^* must be lower than the equilibrium innovation rate when there are no bribes η_1^*. This leads to my main result: The effects of penalties to corruption on equilibrium innovation rates are highly nonlinear. In particular, small changes in penalties $p + m$ around the critical value $(1 - \varphi)/\varphi$ can lead to large changes in innovation rates, while changes in penalties far from that critical value have no effects. The discontinuous dependence of the equilibrium innovation rate η^* on the total penalties $p + m$ is depicted in figure 6.

The effects on the equilibrium innovation rate of changes in the probability of detecting corruption φ and in the fraction of entry applications that get inspected γ are more complex, since they not only determine whether bribes are paid, but also affect the position of

FIGURE 4

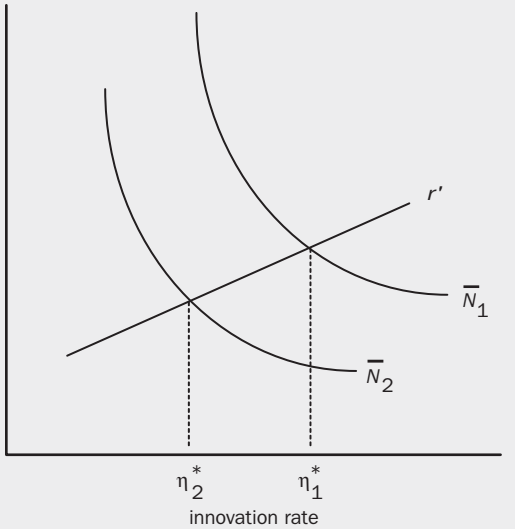
Industry equilibrium



Note: See the text for further details.

FIGURE 5

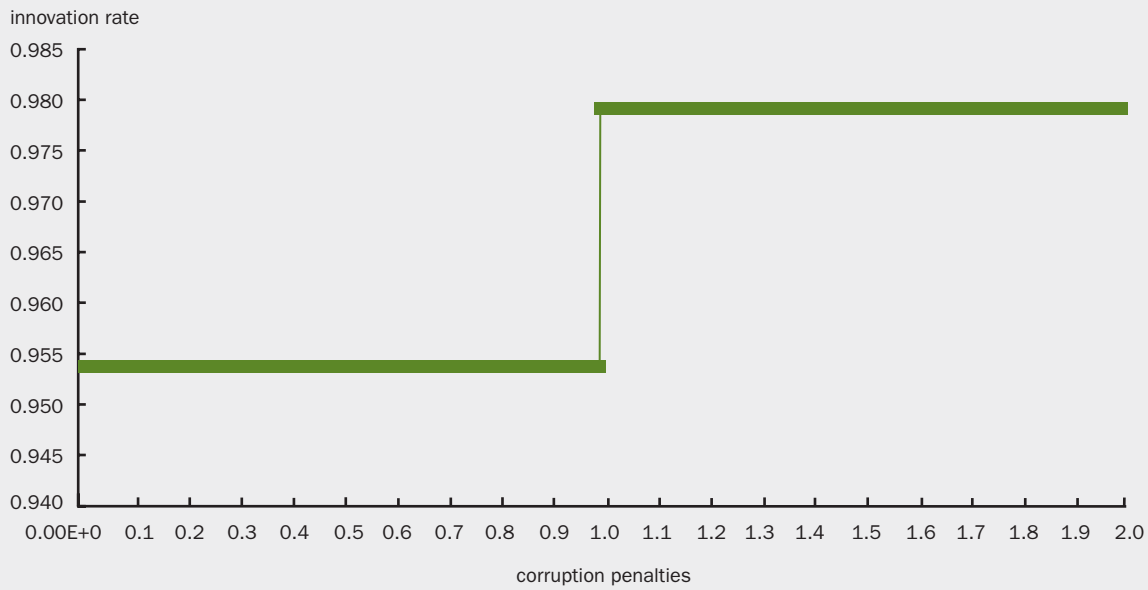
Innovation rates with (η_2^*) and without (η_1^*) bribes



Note: See the text for further details.

FIGURE 6

Innovation rate (η^*) vs. corruption penalties ($p + m$)



the curves $\bar{N}_1$ and $\bar{N}_2$ in figure 5. A numerical analysis of these effects is provided in the appendix.

Conclusion

I have illustrated how the rate of product innovation can be affected by changes in parameter values determining the amount of corruption in an industry. An interesting result of the analysis is that, under certain parameter ranges, small increases in the penalties to corruption or the effectiveness of detection can result in large increases in the amount of product innovation.

While I have not explicitly analyzed the effects of innovation on economic development, it is safe to speculate what those effects would be. To be specific, consider Grossman and Helpman's (1991) endogenous growth model. In that model, there is a continuum of product lines, each characterized by quality ladders of fixed increments. In each product line, there is always a leader producer that supplies the frontier quality and drives all previous producers out of the market. However, the arrival rate of innovators is optimally determined in an R&D sector. Successful innovators drive the incumbent leaders out of the market and become the new product leaders. Thus, each product line has a similar structure as the industry considered in this article. Introducing a corruption game in each product line would thus deliver similar results. Since in Grossman and Helpman (1991) the growth rate of the economy is determined by the endogenous innovation rate, the effects of corruption found here would translate into growth effects. In particular, small increases in the penalties to corruption or the effectiveness of detection can lead to jumps in the growth rate of the economy. Thus, corruption has the potential of grouping countries into two distinct development groups: fast- and slow-growing countries.

NOTES

[1]While I do not explicitly analyze the links between corruption, innovation, and economic growth, I sketch them in some detail in the conclusion.

[2]The Penn World Table—maintained by the Center for International Comparisons at the University of Pennsylvania—provides purchasing power parity and national income accounts converted to international prices for 188 countries for some or all of the years 1950–2004. For further details, please see http://pwt.econ.upenn.edu/. Transparency International is a global organization promoting anticorruption policies. Its Corruption Perception Index ranks countries by the perceived levels of corruption (frequency and/or size of bribes) in the public and political sectors, as determined by expert assessment and business opinion surveys. The Corruption Perception Index can be downloaded from www.transparency.org.

[3]For example, Djankov et al. (2002) report that to meet government requirements for starting a business in 1999, an entrepreneur in Italy needed to follow 16 different procedures, pay US$3,946 in fees, and wait at least 62 business days to acquire the necessary permits. In contrast, an entrepreneur in Canada only needed to follow two procedures, pay US$280, and wait for two days. An extended account of how entry regulation leads to corruption and bureaucratic delays is provided by De Soto (1989). However, he focuses on the Peruvian economy.

[4]Introducing a fine to polluters would significantly complicate the analysis of the corruption game without additional insights.

[5]This bribe request makes the illegal producer indifferent between accepting and rejecting it.

[6]This bribe request makes the legal producer indifferent between accepting and rejecting it.

[7]This cost function is assumed to be increasing, differentiable, and strictly convex. Moreover, $r'(0) = 0$ and $r'(\infty) = \infty$.

APPENDIX: RESEARCH AND DEVELOPMENT DECISIONS AND INDUSTRY EQUILIBRIUM

Given the solution to the corruption game characterized in the main text, we can proceed to write expressions for N_P and N_C. The expected value of an innovator that does not pollute N_C is given by:

$$N_C = \left\{ \begin{array}{c} V \text{ if } p+m > \dfrac{(1+\varphi)}{\varphi} \\ (1-\gamma)V + \gamma\varphi V, \text{ otherwise} \end{array} \right\}.$$

Observe that when $p+m > \dfrac{(1+\varphi)}{\varphi}$, there are no bribes paid in the corruption game. Hence, the clean innovator obtains the value V of becoming a leader with certainty. When $p+m < \dfrac{(1+\varphi)}{\varphi}$, bribes are paid whenever the innovator gets inspected. As a consequence, the innovator gets the full value V only if he is not inspected, an event that happens with probability $(1-\gamma)$. With probability γ, the (clean) innovator is inspected and obtains a value (net of bribes) of φV.

The expected value of an innovator that pollutes N_P is given by:

$$N_P = (1-\gamma)\, V.$$

The innovator that pollutes obtains the full value of becoming the leader V only if he is not inspected, which happens with probability $(1-\gamma)$. With probability γ, the innovator that pollutes is inspected and is precluded from producing (recall that for every parameter specification the government official always takes the legal course of action).

The value of being the industry leader V is given as follows:

$$iV = \left\{ \begin{array}{c} \Pi - \eta V + \eta\xi\gamma V \text{ if } \dfrac{(1-\varphi)}{\varphi} < p+m \\ \Pi - \eta V + \eta\xi\gamma\varphi V, \text{ otherwise} \end{array} \right\},$$

where i is the instantaneous interest rate. The flow value of being the leader iV is given by Π, but with arrival rate η, a new innovator enters the market, in which case the profit flow Π is permanently lost. However, there are exceptions to this loss. When $\dfrac{(1-\varphi)}{\varphi} < p+m$, the loss is avoided when the new arrival pollutes and is inspected by a government official, an event that happens with probability $\xi\gamma$ (in this case there are no bribes imposed and the entry permit is rejected). Also, when $p+m < \dfrac{(1-\varphi)}{\varphi}$, the loss is partly avoided when the new arrival pollutes and is inspected by a government official (again, an event that happens with probability $\xi\gamma$). However, in this case, the leader is only able to retain a fraction φ of the value of being the leader V.

We are now ready to write the expected value of creating a new product generation in equation 4 (p. 33):

$$\bar{N} = \xi N_P + (1-\xi) N_C.$$

This expected value depends on parameter values, since the outcome of the corruption game varies depending on them. As a consequence, I will index the expected value $\bar{N}_j$ according to the parameter region j.

Parameter region $1(j=1)$: $\frac{(1-\varphi)}{\varphi} < p+m$,

A1) $\bar{N}_1(\eta) = \{\xi(1-\gamma)+(1-\xi)\}\frac{\Pi}{i+\eta-\eta\xi\gamma}$.

Parameter region $2(j=2)$: $p+m < \frac{(1-\varphi)}{\varphi}$,

A2) $\bar{N}_2(\eta) = \{\xi(1-\gamma)+(1-\xi)[(1-\gamma)+\gamma\varphi]\}\frac{\Pi}{i+\eta-\eta\xi\gamma\varphi}$.

Observe that, in each parameter region j, the expected value $\bar{N}_j(\eta)$ depends on the industry's arrival rate η, which is an endogenous variable of the model. In particular, the expected values $\bar{N}_j(\eta)$ depend negatively on η. Also, it is straightforward to verify that for every possible value of the arrival rate η, that

A3) $\bar{N}_2(\eta) < \bar{N}_1(\eta)$.

Observe that, since r is a convex function, r' is increasing in η. This, together with the previously mentioned properties for the expected values $\bar{N}_j(\eta)$, allows us to establish that in each parameter region j there is a unique equilibrium arrival rate η_j^* satisfying that

$$r'(\eta_j^*) = \bar{N}_j(\eta_j^*),$$

and that these arrival rates are ordered across parameter regions as follows:

A4) $\eta_2^* < \eta_1^*$.

As mentioned in the main text, this inequality leads to the main result of the article. Fixing all other parameter values, lower penalties on corruption $p + m$ lead to lower rates of innovation. However, the relation is highly nonlinear. Reductions in $p + m$ have no effects on rates of innovation as long as they leave the model within the same parameter region. But once the edge of a parameter region is approached, small reductions in $p + m$ have large effects as the equilibrium innovation rate η^* jumps from one region to the next.

The effects of the probability of detection φ and the fraction of entry applications that get inspected γ are more complex because they affect not only the length of the parameter regions but also the position of the expected values $\bar{N}_1$ and $\bar{N}_2$ in figure 5 (p. 34). To ease the presentation of these effects, in what follows I complement the analysis with a numerical example. It is important to point out that the example has no empirical content, since parameter values are not chosen to reproduce observations; it serves illustration

FIGURE A1

Innovation rate (η^*) vs. probability of detecting corruption (φ)

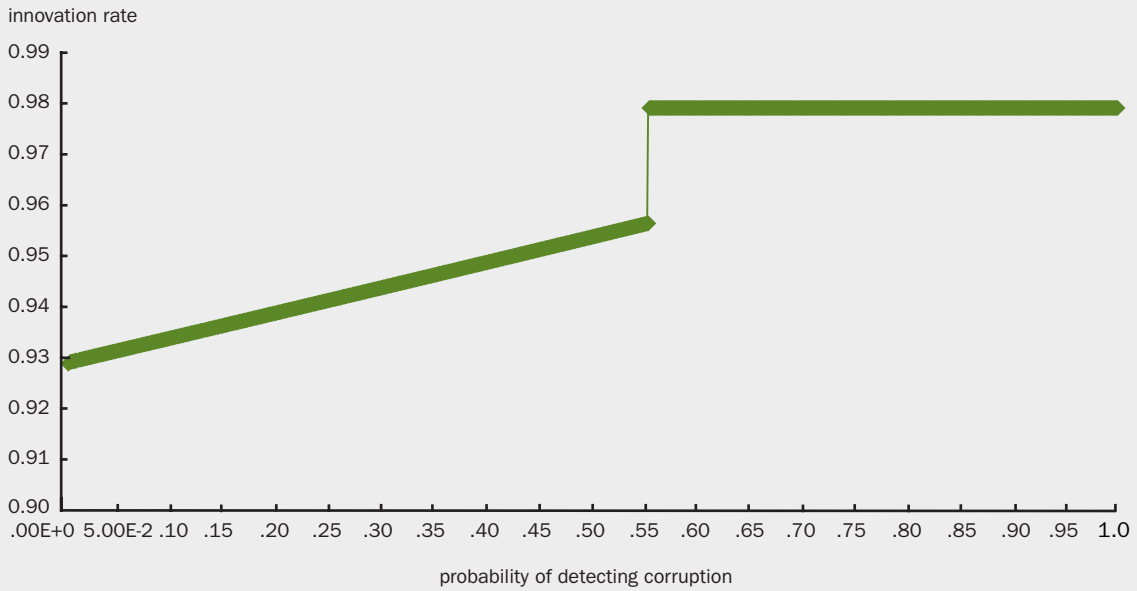
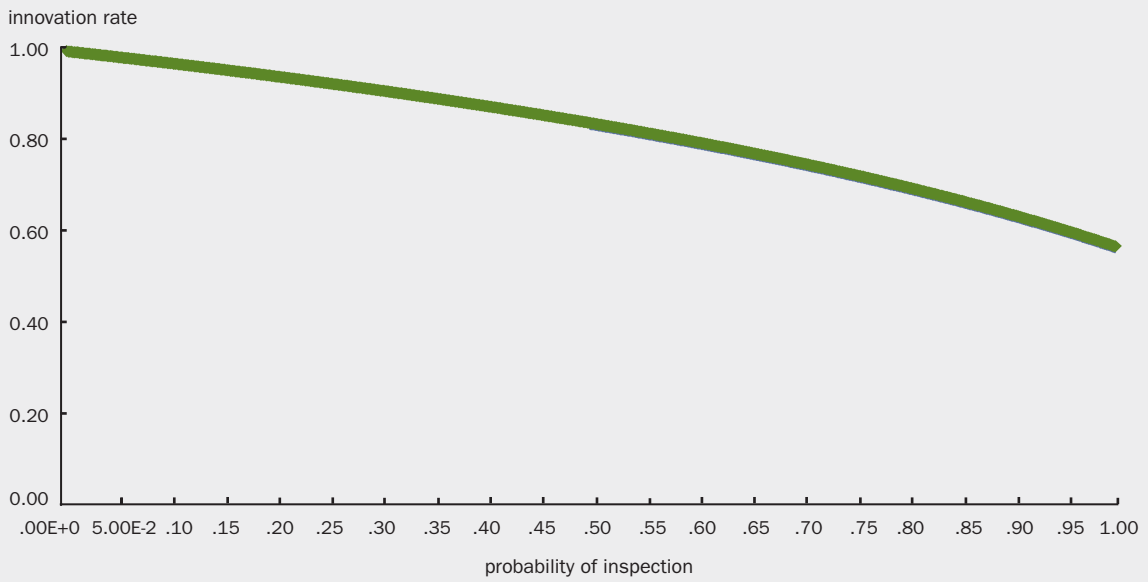


FIGURE A2

Innovation rate (η^*) vs. probability of inspection (γ)



purposes only. The example considered has the following parameter values: $\xi = 0.5$, $\gamma = 0.1$, $\varphi = 0.5$, $p = 0.8$, $m = 0$, $i = 0.04$, $\Pi = 1$ (this is just a normalization), and $r(\eta) = \frac{1}{2}\eta^2$.

Fixing all other parameters at their benchmark values, figure A1 shows how the equilibrium innovation rate depends on the probability of detecting corruption φ. The figure shows that a higher detection probability φ (weakly) increases the innovation rate of the industry. However, the dependence is discontinuous, and once the arrival rate jumps, it is unresponsive to further increases in φ. These properties are general. We see from equations A1 and A2 that $\bar{N}_j(\eta)$ increases with φ when $j = 2$ but is independent of φ when $j = 1$. Moreover, an increase in φ can bring the economy from parameter region $j = 2$ to $j = 1$, entailing a jump in the arrival rate from η_2^* to η_1^* at the critical value for φ at which

$$p + m = \frac{(1-\varphi)}{\varphi}.$$

Figure A2 shows how the equilibrium innovation rate depends on the probability of inspection γ. The figure shows that a higher probability of inspection γ decreases the innovation rate of the industry in a continuous way. This is a general result. We see from equations A1 and A2 that $\bar{N}_j(\eta)$ decreases with γ in each case $j = 1, 2$. Since the functions depicted in figure 5 (p. 34) shift down as γ increases, the intersections with $r'(\eta)$ take place at lower values of η_j^*, for each $j = 1, 2$. However, changes in γ have no effect on the parameter region that the economy lies on. Thus, while the innovation rate decreases with γ, there are no points of discontinuity.

REFERENCES

Acemoglu, D., and T. Verdier, 2000, "The choice between market failures and corruption," *American Economic Review*, Vol. 90, No. 1, March, pp. 194–211.

Becker, G., and G. Stigler, 1974, "Law enforcement, malfeasance, and the compensation of enforcers," *Journal of Legal Studies*, Vol. 3, No. 1, January, pp. 1–18.

De Soto, H., 1989, *The Other Path: The Invisible Revolution in the Third World*, New York: Harper and Row.

Djankov, S., R. La Porta, F. Lopez-de-Silanes, and A. Shleifer, 2002, "The regulation of entry," *Quarterly Journal of Economics*, Vol. 117, No. 1, February, pp. 1–37.

Grossman, G., and E. Helpman, 1991, "Quality ladders in the theory of economic growth," *Review of Economic Studies*, Vol. 58, No. 1, January, pp. 43–61.

Mauro, P., 1995, "Corruption and growth," *Quarterly Journal of Economics*, Vol. 110, No. 3, August, pp. 681–712.

Shleifer, A., and R. Vishny, 1993, "Corruption," *Quarterly Journal of Economics*, Vol. 108, No. 3, August, pp. 599–617.