The Effect of the Social Security Earnings Test on Male Labor Supply

New Evidence from Survey and Administrative Data

Steven J. Haider
David S. Loughran

ABSTRACT

Despite numerous empirical studies, there is surprisingly little agreement about whether the Social Security earnings test affects male labor supply. In this paper, we provide a comprehensive analysis of the labor supply effects of the earnings test using longitudinal administrative earnings data and more commonly used survey data. We find that the response to the earnings test in survey data is obfuscated by measurement error and labor market rigidities. Accounting for these factors, our results suggest a consistent and substantial response to the earnings test, especially for younger men.

I. Introduction

The retirement earnings test is a provision of the Social Security system that reduces the benefits of current beneficiaries who earn above a specified threshold. In 2004, the provision reduced current Social Security benefits by $1 for every $2 earned above $11,640 for individuals who claim benefits before their full retirement age (aged 65 and four months for those turning 65 in 2004). Although

Steven J. Haider is a professor of economics at Michigan State University. David S. Loughran is a senior economist at the RAND Corporation. *The authors gratefully acknowledge the financial support of the Social Security Administration through the Michigan Retirement Research Center and from the National Institute on Aging under grant 5 P01 AG022481-04 and the helpful comments of Jeff Biddle, Dan Black, John Bound, Tom DeLeire, Stacy Dickert-Conlin, Alan Gustman, Kathleen McGarry, Paul Menchik, Deborah Reed, Gary Solon, Mel Stephens, and Steve Woodbury. The findings and conclusions expressed herein are solely those of the authors and do not represent the views of the Michigan Retirement Research Center, the Social Security Administration, the National Institute on Aging, or any other agency of the Federal government.* *The data used in this article can be obtained beginning August 2008 through July 2011 from Steven Haider, 110 Marshall-Adams Hall; Department of Economics; Michigan State University; East Lansing, MI 48824; <haider@msu.edu>.*

[Submitted July 2006; accepted March 2007]
ISSN 022-166X E-ISSN 1548-8004

these lost benefits lead to higher future benefits, popular opinion has long viewed the earnings test as an unfair tax on the earnings of older workers that dramatically reduces their incentive to work. Echoing these concerns, Congress voted to eliminate the earnings test for workers aged 70–71 in 1983 and for workers between the full retirement age and age 69 in 2000.

Understanding how the earnings test affects labor supply is important for at least two reasons. First, the earnings test still covers individuals aged 62 to the full retirement age, and the full retirement age is currently scheduled to increase to age 67 by 2022. Thus, understanding how labor supply responds to the earnings test will be helpful in evaluating any future reform proposals. Second, the earnings test provides us with an opportunity to examine how the elderly respond to taxes on earnings and, more generally, to changes in wages. Although interpreting the labor supply response is complicated by the fact that the taxed benefits lead to higher future benefits, the nature of the response can still provide us with some insight about the extent to which older men are forward-looking and can adjust their earnings.

Much of the earlier empirical literature on the earnings test concludes that the earnings test has little meaningful effect on the labor supply of older men (for example, Viscusi 1979; Burtless and Moffitt 1985; Vroman, 1985; Honig and Reimers 1989; Leonesio 1990; Gustman and Steinmeier 1991; Gruber and Orszag 2003).[1] This conclusion is supported by analyses of the degree to which workers "bunch" at the earnings test threshold, structural models of labor supply that exploit the kinked budget set created by the earning test, and regression analyses of whether the employment and earnings of older men responded to various reforms of the earnings test during the 1980s and 1990s. While the bunching analyses show that a disproportionate number of older male workers have earnings near the earnings test threshold (Burtless and Moffitt 1985; Friedberg 2000), most researchers have concluded that the number of workers who are constrained by the earnings test is small (Leonesio 1990; Gruber and Orszag 2003) and that the response of these constrained workers to changes in the earnings test is likely to be small (Burtless and Moffitt 1985, Gustman and Steinmeier 1985, Honig and Reimers 1989). Analyses of how workers responded to the repeal of the earning test for 70–71-year-olds in 1983 and to changes in the earnings test threshold during the 1980s and 1990s find those reforms had little effect on aggregate male labor supply (Gruber and Orszag 2003).

In contrast, Friedberg (2000) concludes that the earnings test has significant effects on male labor supply using a structural model that exploits both the kinked budget set created by the earnings test and reforms to the earnings test in the 1980s and 1990s. In addition, a number of recent studies have found that the 2000 elimination of the earnings test for 65–69-year-olds raised the employment and earnings of affected male workers (Tran 2003; Song 2004). This finding is consistent with studies of the elimination of similar earnings tests in Canada (Baker and Benjamin 1999) and the United Kingdom (Disney and Smith 2002).

In this paper, we seek to provide a comprehensive view of how men respond to the earnings test by conducting analyses with commonly used survey data from the Current Population Survey (CPS) and with longitudinal administrative earnings data.

1. Several recent textbooks draw the same conclusion. See Borjas (2005, pp. 85–88) and Kaufman and Hotchkiss (2006, pp. 133–35).

The main contributions of these analyses are three-fold. First, we use two sources of administrative data to analyze the extent to which bunching in self-reported earnings data from the CPS accurately depicts whether men respond to the earnings test. We find that, because of measurement error and rigidities in the choice of hours, bunching in self-reported data substantially underestimates the number of workers who respond to the earnings test. Second, we use panel data to show that the labor supply of men aging past the earnings test responds in a manner that is consistent with theory. Third, we examine the labor supply response to the 1983 and 2000 eliminations of the earnings test, documenting both the aggregate response and the response by age. Our results indicate that the labor supply of only relatively young workers responded to the elimination of the earnings test in 2000, which can explain why the 1983 repeal for 70–71-year-olds had no aggregate effect on employment and earnings. Overall, the evidence we report here and evidence reported in previous studies suggest that men respond to the earnings test in a manner that is consistent with theory and relevant for policy.

II. Background

A. *Rules Surrounding the Retirement Earnings Test*

The retirement earnings test is a provision of the Social Security system that reduces the benefits of current Social Security recipients for each dollar earned above a given threshold. The ages covered by the earnings test, the level of the threshold, and the rate at which benefits are reduced have varied considerably over the last three decades. Table 1 shows how these various provisions of the earnings test varied between 1975 and 2004. All beneficiaries aged 62–71 faced the same earnings threshold between 1975 and 1977, and then in 1978, the threshold was increased more for individuals aged 65–71 compared to individuals aged 62–64. The earnings test was eliminated for beneficiaries aged 70–71 in 1983 and for beneficiaries older than the full retirement age in 2000.[2] In nominal terms, the threshold increased steadily over the entire period, with especially large percentage increases in 1978 and the late 1990s. The earnings test reduced Social Security benefits $1 for every $2 earned above the threshold during most of our sample period. In 1990, the rate of benefit reduction was lowered to $1 for every $3 earned above the threshold for beneficiaries aged 65–69. Most labor earnings count toward the retirement earnings test, not just Social Security covered earnings.

There exist several additional provisions of the Social Security system that interact with the retirement earnings test in important ways. The first set of provisions increase future benefits when current benefits are lost due to the earnings test. For workers between age 62 and the full retirement age, the increase in future benefits is computed using the Actuarial Reduction Factor (ARF). When the earnings test covered individuals above the full retirement age, the increase in future benefits

2. Until 2000, the rules were applied according to one's age in each month. Starting in 2000, a different threshold is specified for the year in which someone reaches the full retirement age. This alternative threshold is much higher (for example, $17,000 in 2000, $25,000 in 2001, and $30,000 in 2002).

Table 1
The Social Security Retirement Earnings Test Threshold by Year and Age

	Age		
Year	62–65 (FRA)	65 (FRA)–69	70–71
1975	2,520	2,520	2,520
1976	2,760	2,760	2,760
1977	3,000	3,000	3,000
1978	3,240	4,000	4,000
1979	3,480	4,500	4,500
1980	3,720	5,000	5,000
1981	4,080	5,500	5,500
1982	4,440	6,000	6,000
1983	4,920	6,600	—
1984	5,160	6,960	—
1985	5,400	7,320	—
1986	5,760	7,800	—
1987	6,000	8,160	—
1988	6,120	8,400	—
1989	6,480	8,880	—
1990	6,840	9,360[a]	—
1991	7,080	9,720[a]	—
1992	7,440	10,200[a]	—
1993	7,680	10,560[a]	—
1994	8,040	11,160[a]	—
1995	8,160	11,280[a]	—
1996	8,280	12,500[a]	—
1997	8,640	13,500[a]	—
1998	9,120	14,500[a]	—
1999	9,600	15,500[a]	—
2000	10,080	—[b]	—
2001	10,680	—[b]	—
2002	11,280	—[b]	—
2003	11,520	—[b]	—
2004	11,640	—[b]	—

Notes: All figures are in current dollars. The dashed-lines denote major changes to the retirement earnings test. Unless otherwise noted, the loss in benefits is $1 for every $2 earned above the threshold. Technically, the age cut-off between the first and second groups is the full retirement age (FRA). The FRA was 65 for individuals reaching age 62 from 1975 through 2000 but then increased for individuals reaching age 62 thereafter. For those reaching age 62 in 2004, the FRA was 65 and 10 months.
a. The loss of benefits is $1 for every $3 earned for this age group.
b. Individuals are subject to a different threshold in the year in which they reach the full retirement age (for example, $17,000 in 2000, $25,000 in 2001, $30,000 in 2002) and lower benefits tax ($1 for every $3 of labor earnings).

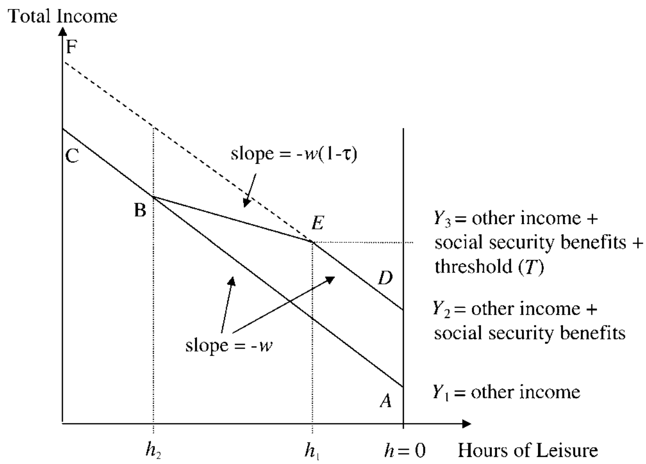

Figure 1
The Retirement Earnings Test and the Budget Constraint

was determined by the Delayed Retired Credit (DRC). The rate at which the ARF increases future benefits is 8 percent per annum, a rate that is believed to be approximately actuarially fair. The rate at which the DRC increases future benefits increased from 3 to 7 percent per annum over our sample period, implying the rate went from being largely actuarially unfair to almost fair.[3] However, many researchers have argued that the vast majority of beneficiaries, their financial advisors, and the media ignore these future benefit increases, and thus, the earnings test is perceived to be a pure tax.

B. Theoretical Predictions Regarding the Earnings Test

Following previous studies, we first consider a standard one-period model of labor supply in which individuals are offered a wage and then choose the number of hours they work.[4] A worker subject to the earnings test who does not claim benefits faces a standard budget constraint that is determined by other income Y_1 and the wage rate w, denoted as line segment AC in Figure 1. For an individual who claims Social Security benefits, the retirement earnings test induces kinks in the budget constraint because the benefit reduction is equivalent to a tax on earnings when the earnings

3. In 1983, a law was passed to raise the DRC from 3–8 percent; the increase was phased in from the 1925/26 birth cohort to the 1943/44 birth cohort. See SSA (2004) for a detailed discussion of the ARF and DRC and its historical changes. See Leonesio (1990) for a detailed discussion regarding actuarial fairness. See Pingle (2006) for a detailed discussion regarding recent changes to the DRC.
4. Similar presentations can be found in Burkhauser and Turner (1978), Blinder, Gordon, and Wise (1980), Friedberg (2000), Borjas (2005), and many other published treatments of Social Security and labor supply. Throughout this discussion, we ignore all other taxes and transfers to focus our attention on the impact of the earnings test.

test binds. More specifically, we denote the sum of other income and Social Security benefits as Y_2, the threshold as T, the benefit reduction rate as τ the quantity of hours where an individual's earnings is equal to the threshold as h_1 ($h_1 = T/w$), and the quantity of hours where benefits are completely taxed away as h_2. Then, the full budget constraint for an individual who claims Social Security benefits and is covered by an earnings test is denoted by the line segment DEBC in Figure 1. If the earnings test were eliminated, then the wage would be w once again and other income would be Y_2, resulting in the budget constraint denoted as line segment DF.

With the earnings test in place, the model predicts that workers will disproportionately "bunch" at the convex kink in the budget constraint (point *E*) because that point is consistent with a range of indifference curves. If the earnings test is eliminated, the simple model predicts that responses should vary across individuals. Some individuals should not respond because the budget constraint does not change (those who would have located on segment *DE*); others should increase hours because the budget constraint pivots inducing a substitution effect (those who would have located at point *E*); others should decrease hours because the budget constraint shifts out inducing an income effect (those who would have located on segment *BC*); and still others should respond ambiguously because the budget constraint pivots and shifts (those who would have located on segment *EB*). Thus, the aggregate response to the elimination of the earnings test is an average over individuals in these four regions of the budget constraint, implying that the effect of the earnings test on labor supply is ambiguous.

Behavior might deviate from the predictions of this one-period model for a number of reasons. One is that individuals may be forward-looking, perhaps behaving in a manner consistent with the life-cycle labor supply model (Burkhauser and Turner 1978). Such a model predicts that individuals will take into account that earning above the threshold leads to higher future benefits through the ARF/DRC provisions, leading individuals to respond less or not at all to the earnings test. For workers above the full retirement age, the DRC is not actuarially fair for most of our sample period, so we would still expect to see some bunching for covered workers above the full retirement age. For covered workers below the full retirement age, however, the ARF is approximately actuarially fair, so we would expect to see less or even no bunching. The life-cycle model also predicts that individuals intertemporally substitute labor supply away from periods of high-taxes to periods of low-taxes. For example, individuals may increase their hours in periods before the earnings test applies and reduce their hours until after the earnings test expires. We would expect the measured change in earnings between high- and low-tax periods to be greater than in a one-period model (Blundell and MaCurdy 1999).

A second reason why behavior may not conform to the predictions of the one-period model is that we have assumed workers can freely choose hours conditional on wages. However, a number of studies argue that labor market rigidities prevent workers from freely choosing hours (Gustman and Steinmeier 1983; Lundberg 1985; Hurd 1996; Rust and Phelan 1997). If older workers cannot choose hours freely, they may not be able to choose hours precisely at the convex kink, making bunching at the threshold less distinct even if individuals perceive the earnings test as a tax. Moreover, when the earnings test is eliminated, we can no longer make clear predictions about how labor supply will respond if hours could not be freely chosen.

For example, individuals who were induced to leave the labor force might return following the elimination of the earnings test (Reimers and Honig 1993).

Labor market rigidities might also affect whether older individuals can freely exit and enter the labor market. Such rigidities could arise if entry and exit costs are high or if human capital depreciates quickly. For example, if the cost of reentering the labor market increases with time out of the labor force, then the effect of eliminating the earnings test might be smaller for targeted individuals who are older.

C. *Implications for Empirical Analyses*

A variety of methods has been used to examine the effect of the earnings test on labor supply, none of which is completely satisfactory. Bunching analyses examine whether a disproportionate number of workers have earnings in the vicinity of the earnings test threshold. This bunching provides an estimate of the number of individuals who are constrained by the earnings test (Leonesio 1990; Friedberg 2000) and can be used to identify behavioral parameters (Burtless and Moffitt 1985; Friedberg 2000; Saez 2002). However, bunching analyses typically do not account for measurement error or labor market rigidities, both of which cause observed bunching to understate the actual labor supply response.

Aggregate labor supply responses to changes in the earnings test could also understate the degree to which workers are constrained. Theory suggests that the direction of the labor supply response depends on an individual's initial location on the budget constraint. Consequently, a small aggregate response could represent small but consistent labor supply responses or large but offsetting labor supply responses.

To provide a comprehensive assessment of how the earnings test affects male labor supply, we use 30 years of CPS data to examine bunching and aggregate labor supply responses, similar to previous studies. We supplement those analyses with a direct assessment of the role of measurement error and rigidities using administrative data, new longitudinal evidence on how men respond to aging past the earnings test, and parallel estimates of the aggregate labor supply response to the 1983 and 2000 repeals of the earnings test. Taken together, we argue that a consistent picture of how men respond to the earnings test emerges from these various analyses.

III. Data

Our analyses use three different data sets, each with their own benefits and drawbacks. We provide an overview of the three data sets in Table 2 and summary statistics in Appendix tables.

A. *Current Population Survey (CPS) March Demographic Files*

We use data from the March CPS for the years 1976 through 2005. Each CPS survey provides earnings information for the previous year, implying we have earnings data for the years 1975 through 2004. Throughout the rest of the paper, we will refer to the data by the year for which earnings are reported and define age to be one year younger than the reported age in the survey year. We restrict our CPS sample

Table 2
Sample Characteristics

	CPS	NBDS	BEPUF
Years covered	1975–2004	1983–1989	1990–2004
Panel	No	Yes	Yes
Persons	186,340	4,769	145,054
Person-years	—	31,788	1,098,664
Source of earnings data	Self-report	Administrative	Administrative

Notes: Sample sizes include all men aged 63–76 for the years reported.

to individuals aged 63–76 during the year of reported earnings. We measure total labor earnings in the CPS as the sum of wage and salary, self-employment, and farm earnings. Earnings in the CPS are self-reported by a household respondent. Because the CPS is a stratified random sample, we use the CPS-provided weights for all of our analyses.

The benefit of using the CPS is two-fold. First, it allows us to make comparisons over a large time period with a large, consistent data set. Second, because many previous studies have used the CPS, it allows us to provide direct comparisons to previous results and analyze the measurement error that is contained in a commonly used data set.

Using the CPS also has several important drawbacks. The most important is that earnings are self-reported and thus are subject to reporting error. A second drawback is that, although current receipt of Social Security benefits is recorded, eligibility is not. Individuals who are not eligible for Social Security should not respond to the earnings test and their inclusion in our sample will cause us to understate the effect of the earnings test on the eligible population.[5] A final drawback is that the CPS offers very limited panel information and the utilization of the panel information is subject to additional measurement error.

B. New Beneficiary Data System (NBDS)

The NBDS is a sample of Social Security beneficiaries who first received benefits between mid-1980 and mid-1981. The NBDS, conducted by the Social Security Administration (SSA), first interviewed respondents in 1982 and then interviewed them again in 1991. The data contain extensive information on respondents' demographic characteristics, labor supply, health, household income, and wealth. The data also include matched administrative records of covered earnings from 1951–99. Our NBDS sample, when weighted, is nationally representative of men who first received retired

5. SSA estimates that the fraction of men and women who do not receive benefits as a retired worker, the spouse of a retired worker, or the survivor of a retired worker is less than 8 percent. Therefore, any understatement of program effects should be minor (SSA 2004).

worker benefits during the sample selection period, qualified for benefits based on their own earnings history, and did not receive Social Security Disability Insurance payments before they retired. We follow these men between the ages of 63–76 over the years 1983–89 for each full calendar year they are alive.[6]

The primary benefit of the NBDS is the matched earnings records. These data allow us to examine the earnings of individuals who are subject to the Social Security earnings test relatively error free and over time. In particular, it is exactly the earnings reported to SSA that matter for the earnings test, and because the administrative earnings data are reported to four digits of significance, we observe Social Security earnings to the dollar in the neighborhood of the earnings test. Because we only use the NBDS to examine bunching near the threshold and the threshold is much lower than the Social Security taxable limit, the fact that the earnings are censored at the taxable limit is not problematic. The primary drawback for our purposes is that the NBDS sampling frame makes it representative of a different population than the population that the CPS represents. We make a number of sample restrictions (detailed below) to both the NBDS and CPS in order to minimize differences between their sampling frames.

C. 2004 Social Security Benefit and Earnings Public Use File (BEPUF)

The BEPUF is a nationally representative data set of Old-Age, Survivors, and Disability Insurance (OASDI) beneficiaries who were entitled to receive OASDI benefits in December 2004. The data only include information available in Social Security administrative records, such as age, gender, benefit level, type of benefits received, and annual Social Security covered earnings from 1951 through 2003.

Like the NBDS, the BEPUF provides administrative data and panel data on earnings. The BEPUF has several additional advantages. First, the BEPUF provides very large sample sizes (see Table 2). Second, the BEPUF allows us to examine a more recent time period, including the 2000 elimination of the earnings test. Third and perhaps most important, the sampling frame of the BEPUF is much more comparable to the CPS, especially for older men.[7] The drawback to the BEPUF is that it contains relatively little information about individuals. In addition, our analysis of the 2000 elimination of the earnings test with the BEPUF is affected by the fact that BEPUF earnings are censored at the taxable limit.

6. We make two additional refinements to our sample. First, we restrict our sample to individuals who were born after 1912, so that we observe each sample member at an age younger than 70-years-old at least once during our sample period. Second, we include 63-and 64-year-olds in 1982, so that we have a larger sample of younger individuals.

7. The BEPUF is nationally representative of all current beneficiaries of Social Security, not just new beneficiaries as in the NBDS. Thus, for those ages in 2004 in which most eligible men will have claimed benefits, the BEPUF will be nationally representative for everyone that is eligible for benefits. Tabulations from the BEPUF indicate that, among eligible individuals who will eventually claim benefits, more than 97 percent claim by age 68 and more than 99 percent claim by age 70. In contrast, because the NBDS sampled new beneficiaries, it is largely composed of individuals who were aged 62 and 65 in 1980–81. Although the NBDS sampling frame is not generally representative of any particular age group, it is unclear how the difference in its sampling frame would affect our analysis.

IV. Bunching Near the Earnings Test Threshold

In this section, we use the CPS to show how the magnitude of bunching near the earnings test threshold changed between 1975 and 2004. This analysis extends Friedberg (2000) by providing new evidence of how bunching changed over the last decade. We then use administrative data to examine the extent to which self-reported earnings in the CPS obfuscates the extent to which workers bunch near the threshold.

A. Bunching over Time in the CPS

To examine bunching, we present histograms that plot the fraction of workers reporting earnings within a certain percentage band of the relevant earnings threshold. Unless otherwise noted, we divide each histogram into bins representing increments of ten percentage points relative to a given threshold. For example, a bin labeled "–100" contains individuals who have earnings between 100 percent below and 90 percent below a given threshold, and a bin labeled "–90" contains individuals who have earnings between 90 and 80 percent below the threshold.[8]

Figure 2 presents a panel for each of five time periods (1975–77, 1978–1982, 1983–1989, 1990–1999, and 2000–2004) using the full CPS sample. Each panel presents a histogram separately for the 66–69 age group, 70–71 age group, and 72–74 age group. The bins (along the x-axis) are 10 percentage point bins with respect to the earnings threshold for 65–69-year-olds. Although the histograms only show the bins for positive earnings within 100 percent of the earnings threshold, the *y*-axis shows the percent of all individuals within the age group who fall in each of the bins. Analyses below use alternative denominators to construct the histograms.[9]

Workers bunch exactly as the simple theory predicts under each earnings test regime. In the 1975–77 regime (Panel A) and the 1978–82 regime (Panel B), the youngest two age groups bunch just below the kink, and the oldest age group does not appear to bunch at all. This pattern is consistent with retirement earnings test only covering the first two age groups (with the same threshold) and not covering the older age group. The bunching behavior changes during the 1983–89 regime (Panel C) with the 70–71-year-olds behaving like the 72–76-year-olds, consistent with the elimination of the earnings test for 70–71-year-olds in 1983. The younger 65–69 age group continues to bunch at the threshold. The 1990–99 regime in Panel D is similar to the 1983–89 regime, although the bunching for the 65–69-year-olds is less pronounced. The general decline in bunching across panels is consistent with three policy changes over time: an increase in the real value of the threshold, a decrease in the benefit reduction for excess earnings in 1990 ($1 of reduced benefits for each $3 of excess earnings rather than the previous rate of $1 for each $2 of excess

8. More precisely, let E_i^* be the earnings of individual i divided by the specified threshold and let $\{\gamma_b\}_{b=1}^{B}$ be a sequence of bin starting values. We consider individual i to be in bin γ_b if $\gamma_b < E_i^* \times 100 \leq \gamma_{b+1}$.
9. Specifically, subsequent analyses use as the denominator the number of individuals who are in any of the graphed bins. To make the specific choice of a denominator irrelevant, our analyses of all histograms focus on the ratio of percentages between two bins.

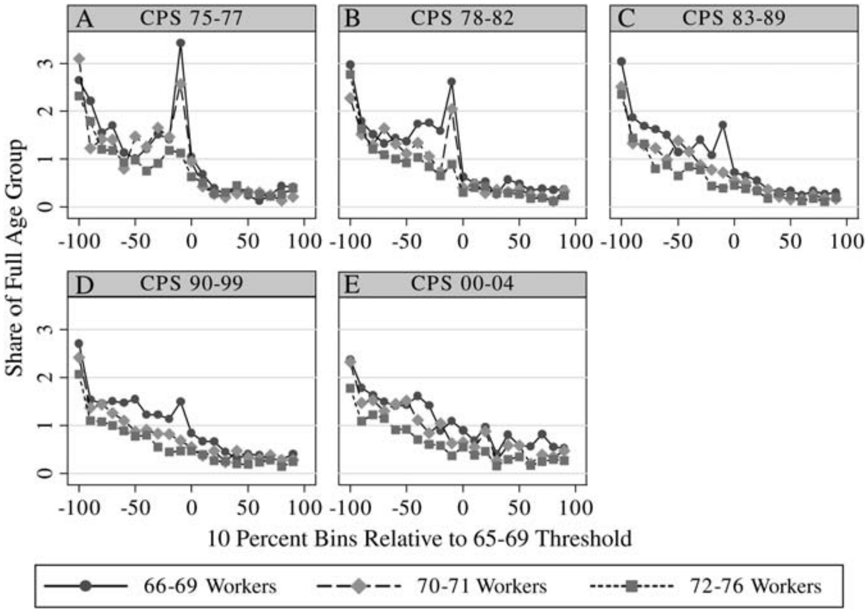

Figure 2
Bunching Relative to the 65–69 Earnings Threshold, 1975–2004

earnings), and an increase in the DRC for some cohorts.[10] Panel E shows that bunching is no longer evident during 2000–2004 among the 66–69 age group, as would be expected with the elimination of the earnings test for 66–69-year-olds in 2000.[11]

Figure 3 presents a similar set of results for 63–64-year-olds and 66–69-year-olds but defines bins relative to the threshold for 62–65-year-olds. Again, workers bunch exactly as the simple theory predicts. The 63–64-year-olds bunch at the same earnings level as the 66–69-year-olds during 1975–78 when they face the same earnings test (Panel A). During the next four earnings test regimes, there remains evidence of bunching for the 63–64-year-olds just below the threshold, where the bunching behavior moves to higher bins (Panels B, C, and D) and then disappears (Panel E) for the 66–69-year-olds.[12] Once again, the degree of bunching diminishes over time, which

10. See Gustman and Steinmeier (1985, 1991, 2004) for a structural model of retirement that estimates the effect of these and similar proposed changes to the retirement earnings test.
11. In Appendix Table A4, we provide point estimates and confidence intervals regarding the extent of bunching, so that statistical significance can be considered. For example, our estimates suggest that the amount of bunching observed for 66–69 workers in 1990–99 is outside the confidence interval for the amount of bunching for 72–76 workers in 1990–99 and for 66–69 workers in 2000–2003, consistent with the earnings test only affecting the labor supply of the first of these three groups. Although such a presentation of statistical significance ignores the fact that all of the measures of bunching are estimates and that the estimates are potentially correlated, the presentation does allow for numerous different comparisons.
12. The point estimate for the amount of bunching for 63–64 workers in 2000–2003 is outside the confidence interval for the amount of bunching for 66–69 workers in 2000–2003; see Appendix Table A4.

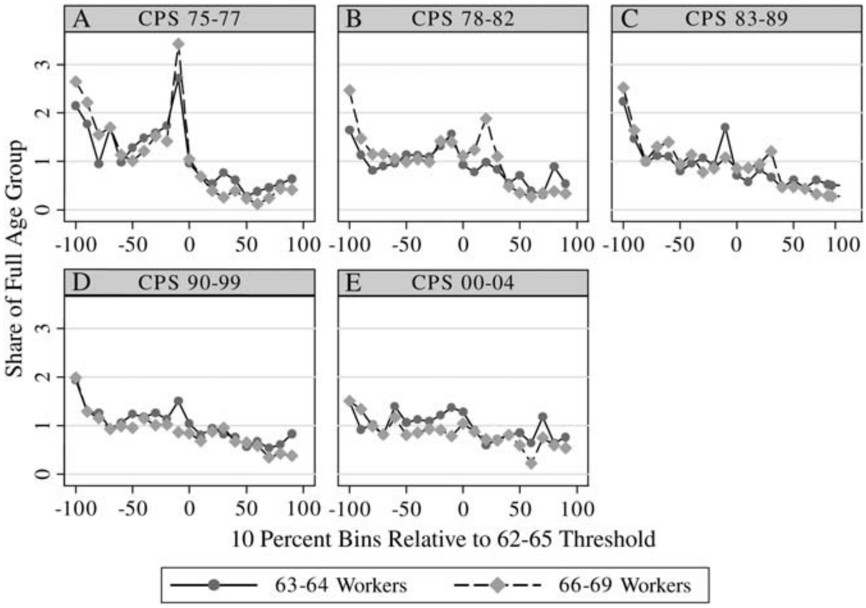

Figure 3
Bunching Relative to the 62–65 Earnings Threshold, 1975–2004

is consistent with the increase in the real value of the threshold. Importantly, these figures suggest that workers aged 63–64 bunch just below the threshold, despite their lost benefits being refunded through the ARF at an approximately actuarially fair rate.

These figures make clear the conclusions from previous studies. Vroman (1985), Friedberg (2000), and many others have concluded that bunching behavior is clearly evident and it moves as the threshold moves. At the same time, Burtless and Moffitt (1985), Leonosio (1990), Gruber and Orszag (2003), and many others dismiss the bunching evidence as being inconsequential. These studies focus on the fact that the number of people bunched at the kink is small in an absolute sense. For example, Figures 2 and 3 suggest that the excess number of people in the bin just below the threshold might be one-half of 1 percent.

B. Assessing the Role of Measurement Error in the CPS with the NBDS and BEPUF

To assess the extent of measurement error in the self-reported CPS earnings, we compare bunching with the administrative earnings data in the NBDS and BEPUF to the self-reported earnings data in the CPS.[13] The benefit of using the NBDS and the

13. Unfortunately, self-reported earnings in the NBDS covers only the previous three months, preventing us from making internal comparisons between the NBDS self-reported and administrative earnings data. No survey data are available for the BEPUF.

BEPUF is that, taken together, the data allow us to examine bunching from 1983 to 2003. For both of these analyses, we assume that the Social Security earnings data represent true earnings, as has been assumed in many previous studies (Bound and Krueger 1992).[14]

We make restrictions on each data set to make them as comparable as possible. For the CPS and the NBDS, we restrict our sample to current Social Security beneficiaries. This restriction helps with comparability because the NBDS only contains information on individuals who are eligible for Social Security benefits, and for the time period we analyze with the NBDS, the NBDS sample must have already applied for benefits. We refer to these samples of current beneficiaries as the ''NBDS-cb'' and ''CPS-cb''. Given the information available in the BEPUF, we cannot select only current beneficiaries, but we can select individuals after they have applied for benefits. We refer to this sample of beneficiaries as the ''BEPUF-b.'' Even with these sample restrictions, however, the data sources are not directly comparable. Whereas the NBDS is effectively a sample of new recipients that are then followed over time, the BEPUF and CPS are effectively samples of current recipients. It is not obvious how these differences in sample design affect our analyses.

We compare the CPS-cb, NBDS-cb, and BEPUF-b in Table 3. Comparing the CPS-cb and NBDS-cb, the average birth year is one year older in the CPS-cb (1918 vs. 1917), consistent with the NBDS being a panel of younger individuals who age over the sample period. The average education (10.8 years) and the percent white (88 percent) are very similar across the two samples. The employment rate is broadly comparable across the two samples, although there is variation year-to-year. Given the limited information available in the BEPUF, we can make even fewer comparisons between the CPS-cb and BEPUF-b. Mean birth year and employment rates are not as close between the CPS-cb and BEPUF-b as were the comparisons to the NBDS-cb.

We examine measurement error in bunching among 63–64-year-olds in Figure 4. To facilitate comparisons across the samples and age groups, we construct the histograms so that the *y*-axis gives the percent of workers locating in a specific bin compared to the total individuals in any bin on the graph.[15] Panels A, B, and C show the amount of bunching in the self-reports of the CPS for three time periods (1983–89, 1990–99, 2000–04); Panels D, E, and F show the amount of bunching in our administrative data for the same time periods. In each panel, we graph the histograms for 63–64-year-olds who are subject to the threshold, and for purposes of comparison, we graph the histograms for 71–74-year-olds who are not subject to any threshold. Each panel shows that the 63–64 age group bunches just below the threshold (the –10 bin), but the 71–74 age group does not. More importantly, the administrative data panels (D, E, and F) exhibit a more pronounced spike just below the threshold

14. In Haider and Loughran (2007), we show evidence on bunching from a fourth data set, the 1978 CPS-Social Security Summary Earnings Exact Match File. The benefit of these data is that they contain self-reported and administrative earnings for the same person. With these data, we find very similar bunching results, suggesting that sample differences are not driving the results we report here. In addition, we show suggestive evidence that the excess bunching in the Social Security earnings records is not due to individuals illegally misreporting earnings to avoid the earnings test.

15. In other words, the denominator for computing percentages is now the total number of individuals in any of the graphed bins. The denominator in previous figures was the total number of individuals in the sample. Again, we only compare the ratio of bins so that the choice of a particular denominator is irrelevant.

Table 3
Comparability of the NBDS-cb, BEPUF-b, and CPS-cb Samples

	1983–89		1990–99		2000–2004	
	CPS-cb	NBDS-cb	CPS-cb	BEPUF-b	CPS-cb	BEPUF-b
Sample size	28,756	25,340	36,597	438,207	20,738	258,354
Birth year						
Mean	1917.7	1917.0	1924.8	1926.5	1932.4	1932.6
Standard deviation	4.0	1.5	4.4	4.1	3.8	3.5
Education						
Mean	10.8	10.8	—	—	—	—
Standard deviation	3.7	3.5	—	—	—	—
White	0.87	0.88	—	—	—	—
Black	0.08	0.08	—	—	—	—
Hispanic	0.03	0.02	—	—	—	—
Working by age						
63	0.26	0.25	0.29	0.39	0.30	0.36
64	0.30	0.24	0.32	0.37	0.37	0.38
65	—	—	—	—	—	—
66	0.27	0.21	0.29	0.40	0.33	0.43
67	0.24	0.21	0.24	0.37	0.32	0.39
68	0.25	0.22	0.24	0.34	0.30	0.36
69	0.24	0.22	0.23	0.32	0.25	0.34
70	—	—	—	—	—	—
71	0.19	0.19	0.19	0.28	0.24	0.29
72	0.19	0.21	0.19	0.26	0.19	0.27
73	0.18	0.20	0.16	0.24	0.18	0.25
74	0.16	0.21	0.15	0.22	0.17	0.23

Notes: See text for definition of CPS-cb, NBDS-cb, and BEPUF-b. Data are only available for BEPUF through 2003; we use 2004 CPS data for sample size considerations.

than do the self-reported data panels (A, B, and C), and the administrative data panels exhibit a steeper decline after the threshold.[16] This latter finding is consistent with the theoretical prediction that workers just above the threshold reduce their hours worked in response to the earnings test.

To quantify the effect of measurement error, we measure the degree of bunching in each graph by comparing the ratio of workers in the –10 bin to the 10 bin.[17] For

16. For each of the three comparisons, the amount of bunching in the administrative data is outside the confidence interval for the amount of bunching in the self-reported data; see Appendix Table A4.
17. We measure bunching by comparing the ratio of workers just below the kink to just above the kink because theory suggests not just that workers should disproportionately locate at the kink, but rather that workers should move from just above the kink to just below the kink.

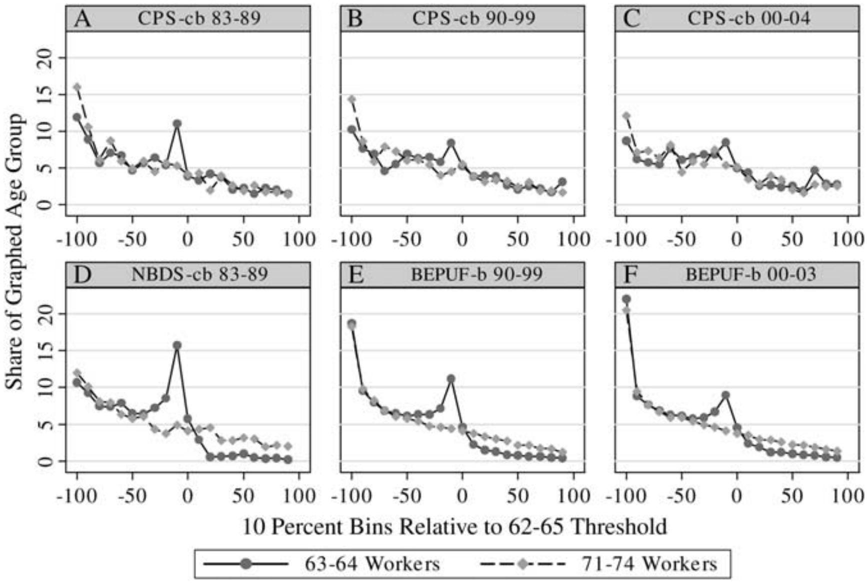

Figure 4
Measurement Error in Bunching for 63–64 Year Olds

example, the ratio for the 1983–89 CPS data is 3.3 (11.0 in the –10 bin and 3.3 in the 10 bin, Panel A), and the ratio for the 1983–89 NBDS data (Panel D) is 5.56. These ratios suggest that bunching in the NBDS administrative data is 69 percent greater than in the comparable CPS self-reported data. Similarly, the amount of bunching in the 1990–99 BEPUF data is 129 percent greater than in the 1990–99 CPS data (a ratio of 2.2 in Panel B versus 5.0 in Panel E), and again, the 2000–2004 BEPUF bunching is 96 percent greater than 2000–2004 CPS bunching (a ratio of 1.9 in Panel C versus 3.8 in Panel F). Thus, in each case, the administrative data exhibits substantially more bunching than does the CPS self-reported data.

We present a similar set of results for 66–69-year-olds in Figure 5. The same general features are apparent, although somewhat less distinct for 1990–99: bunching exists just below the threshold, the amount of bunching is greater with the administrative data, and the decline in workers near the threshold is steeper in the administrative data. The amount of bunching in the administrative data as compared to the CPS self-reports is 64 percent higher for 1983–89 (3.0 in Panel A versus 5.0 in Panel C) and 57 percent higher for 1990–99 (2.5 in Panel B versus 3.9 in Panel D).

Measurement error in the CPS is even more apparent when we examine smaller bins near the threshold. In Figure 6 we show one percent bins (labeled –10, –9, etc.) around the threshold for 66–69-year-olds. In the administrative data, we see that the spike in the fraction of workers with earnings just below the threshold in Figure 5 (based on 10 percentage point bins) is driven almost entirely by workers locating in the –1 bin (workers with earnings between one and zero percent below the

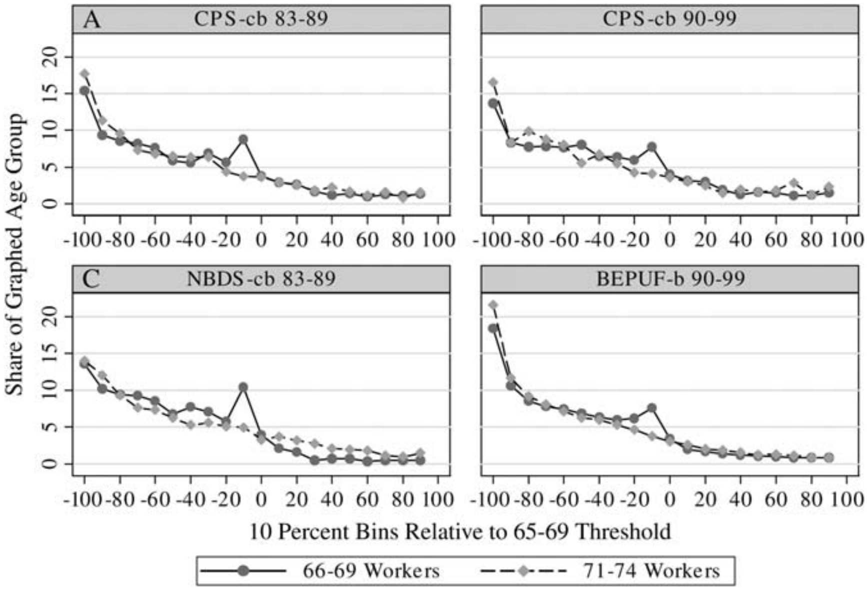

Figure 5
Measurement Error in Bunching for 66–69 Year Olds

threshold). This result demonstrates a remarkable degree of programmatic knowledge and employment flexibility among workers because one percent of the earnings test during the 1980s is about \$80 and \$120 during the 1990s. There is no distinct spike at the –1 bin in the CPS data.

Many studies that examine bunching with survey data acknowledge the potential problem of measurement error. Friedberg (2000) notes that individuals often report earnings to just one or two digits of significance, and thus, she argues that using \$1,000 bins around the earnings test will minimize measurement error problems. Following Friedberg (2000), we also examine the role of measurement error by using \$1,000 bins around the threshold rather than the 10 percentage point bins in Figures 4 and 5 (figures not shown). We find that, for four of the five comparisons, the relative increase in bunching when comparing administrative data to self-reported data is even greater for analyses employing \$1,000 bins than for analyses employing 10 percentage point bins.[18]

Our results based on the NBDS and BEPUF administrative data suggest that measurement error in CPS self-reported earnings obfuscates 60 to 120 percent of the bunching just below the threshold (in the –10 bin). Given that previous studies have

18. Measurement error is worse for the 63–64 age group for all three time periods and for the 66–69 age group for the 1983–89 time period. The intuition for why measurement error is even worse with the \$1,000 bins for these three groups is readily apparent in Figures 4 and 5: the fraction locating in bins above the threshold, the denominator in our calculations, is close to zero.

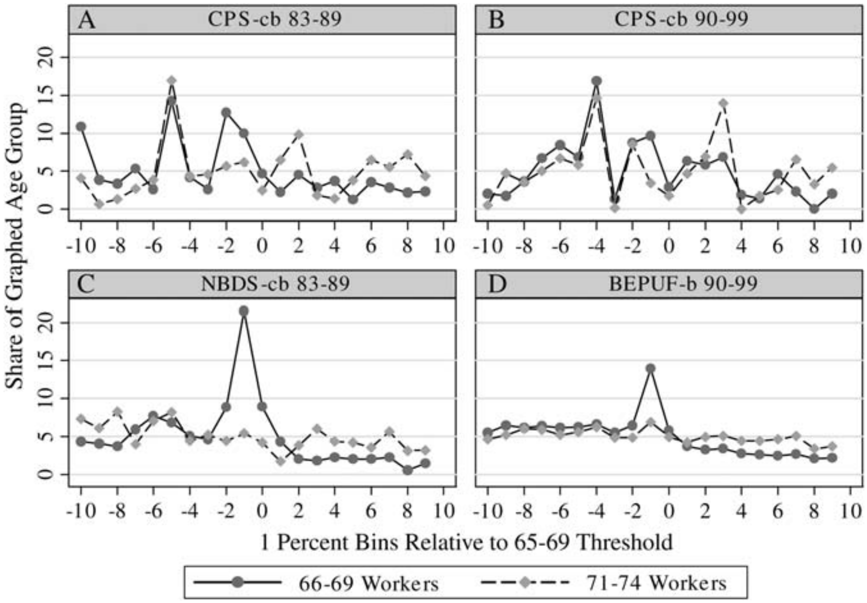

Figure 6
Measurement Error in Bunching for 66–69 Year Olds, 1 Percent Bins

dismissed observed bunching as being inconsequentially small, this finding that measurement error causes bunching to be understated is substantively important.

C. Assessing the Role of Rigidities with the NBDS and BEPUF

Another benefit of the administrative data is that, with less noise, other systematic patterns emerge. Specifically, each panel in Figures 4 and 5 that uses administrative data shows relatively more individuals locating in the bins just below the −10 bin when compared to the older workers (71–74-year-olds). This finding is consistent with the existence of labor market rigidities keeping some workers who are affected by the earnings test from locating precisely at the threshold, despite the fact that a disproportionate share of other workers were locating very close to the threshold (see Figure 6). Importantly, such rigidities imply that even the bunching analysis with the administrative data understates the extent to which individuals are locally changing their labor supply due to the earnings test.

As a rough measure of the effect of rigidities, we compare the excess number of younger workers (63–64-year-olds for Figure 4 and 66–69-year-olds for Figure 5) locating in the −40 to −10 bins to the number of excess workers locating in the −10 bin. We compute excess workers in these bins by subtracting the percent of older workers in these bins from the percent of younger workers. This definition of excess workers is motivated by the relatively close correspondence of the percent of younger and older workers locating in the −100 through −50 bins, suggesting that the relative distribution of workers across bins might be similar if it were not for the earnings test. Despite the close correspondence in the initial bins, our measure should be interpreted with caution because the older workers differ from younger workers along many dimensions that affect labor supply, including health, income, and wealth. We then calculate the ratio of excess workers in the larger number of bins (−40 to −10 bins) to the −10 bin so that the choice of histogram denominator is irrelevant.

Our simple measure of rigidities implies that bunching for 63–64 workers in 1983–89 (Figure 4, Panel D) is underestimated by 74 percent (19 percent excess workers in the −40 to −10 bins versus 11 percent excess workers in the −10 bin). The similar computation for 63–64 workers is 73 percent in 1990–99 (Figure 4, Panel E) and 72 percent in 2000–03 (Figure 4, Panel F). The similar computation for 66–69 workers is 86 percent in 1983–89 (Figure 5, Panel B) and 69 percent in 1990–99 (Figure 5, Panel D). Thus, our simple measure returns remarkably similar estimates across the various time periods and age groups, and it suggests that the existence of rigidities additionally obfuscates the degree of bunching by 70–80 percent.

Two caveats exist about the rigidities analysis. First, the analysis ignores the potential that rigidities cause some workers to locate in bins other than the −40 to −10 bins or even leave the workforce. For example, Figures 4 and 5 provide some evidence of excess workers locating in the 0 and 10 bins. To the extent that rigidities cause workers who respond to the earnings test to locate in bins other than the −40 to −10 bins, our estimates will understate the effect of rigidities. Second, our measure of rigidities uses an arbitrary measure of bunching as its benchmark—the excess workers locating in the −10 bin. Although this benchmark is arbitrary, it is similar to the previous studies that used $1,000 bins because the earnings test is in the neighborhood of $10,000 during much of our sample period.

Despite these caveats, the rigidity results further reinforce the conclusion that the labor supply response to the earnings test is far greater than indicated by bunching in CPS self-reported earnings. As a rough illustration, between 1990 and 1999 the CPS indicates that 1.5 percent of workers aged 63–64 had earnings just below the threshold (the −10 bin in Figure 3, Panel D) compared to 0.8 percent just above the threshold (the 10 bin in Figure 3, Panel D), or the excess workers locating in the −10 bin is 0.7 percent of the population. Our measurement error calculations suggest that the true number of excess workers locating in the −10 bin is about 80 percent greater (0.7×1.8 or 1.3 percent of the population), and our rigidity calculations suggest that the true number of excess workers locating in the −40 to −10 bins is additionally about 75 percent greater (1.26×1.75 or 2.2 percent of the population). For this time period, only about 46 percent of the population was working (see Appendix Table A1). Thus, our calculations suggest that 2.2 percent of the total population or 4.8 percent of the workforce responded to the earnings test. These numbers are far greater than those reported in previous studies (for example, Gruber and Orszag 2003).

V. The Effect of Aging Past The Earnings Test

Over the period of our NBDS and BEPUF samples (1983–99), the earnings test applied to individuals aged 62–69 but not to individuals aged 70 and older. Longitudinal data allow us to examine if and for whom earnings increase as individuals age past the earnings test. The one-period model outlined in Section II predicts that there should be little or no relative change in earnings for those with earnings less than the threshold at age 69, an increase in relative labor supply for those with earnings equal to the threshold at age 69, and an ambiguous response for those with earnings above the threshold at age 69.[19]

To examine the effect of aging past the earnings test, we graph the growth in earnings over a two-year period by the level of initial earnings. The motivation for examining earnings growth over two years is that, during the calendar year in which an individual turns age 70, earnings in months before the month of birth are subject to the earnings test but earnings in subsequent months are not. Thus, to capture the full effect of facing the earnings test, we compare 69-year-olds to 71-year-olds. We account for the secular decline in labor supply with age by comparing two-year changes in earnings for those aging past the earnings test (workers aged 69 in the initial period) to two-year changes in earnings for workers aged 65–67 and aged 71–74 in the initial period. We present results for 1983–89 with the NBDS in Panel A and for 1990–97 with the BEPUF in Panel B of Figure 7.[20]

The figures show that two-year earnings growth is substantially negative, about −20 to −25 percent, across most of the earnings distribution for all three age groups. These large declines in earnings should be expected given the age range of the workers. The exception, however, is workers aged 69 who were earning amounts at or below the threshold. For workers with age-69 earnings near the threshold, earnings did not decline or even increased; workers just below the threshold experienced earnings growth rates about 30 percentage points higher than younger and older workers. Statistical tests indicate that the earnings growth rates for 69-year-olds are significantly greater than the growth rates for the other age groups in the bins at or just below the threshold.[21]

The results for aging past the earnings test reinforce the findings from the bunching analysis. Workers aged 69 at the kink experience substantially higher two-year earnings growth than do younger and older workers at the kink, consistent with the earnings test affecting the labor supply decision of 65–69-year-olds. In addition, the finding that there is also greater earnings growth in bins near the threshold (–40, −20, and 0 bins for the NBDS and −20 through 10 bins for the BEPUF) for 69-year-olds is consistent with rigidities preventing some constrained workers from choosing earnings precisely at the threshold.

19. Because the administrative earnings are censored at the Social Security taxable limit, we are not able to use these data to examine the change in earnings for those at the highest hours. Moreover, this censoring causes all of our results on earnings growth to be downward biased.

20. We graph earnings growth from the -70 to 90 bins. We drop the initial bins because these percentages are quite noisy, presumably because we are dividing by a small earnings level. Because cell sizes become quite small with the NBDS, we only graph up to bin 30 so that we retain at least ten individuals in each bin.

21. To examine whether the differences in earnings growth rates were statistically significant across age groups, we computed a series of two-tailed *t*-tests bin by bin. Given the similarity in earnings changes among the 65–67 year olds and the 71–74-year-olds, we grouped these ages together and tested whether mean two-year earnings changes were different for 69-year-olds. The *t*-tests indicated that the growth among 69 year olds was greater and statistically significant at the 0.05 confidence level for the –40, –20, –10, and 0 bins with the NBDS (Panel A) and for the –20, –10, 0 and 10 bins with the BEPUF (Panel B). In no case was the growth rate lower and statistically significant for 69 year olds.

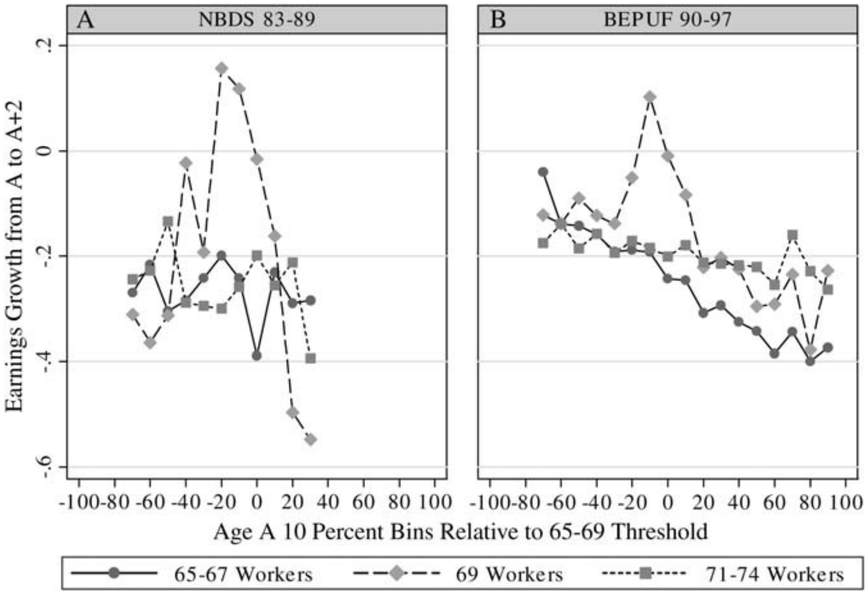

Figure 7
Average Earnings Growth from Age A to A+2 by Age A Bin

VI. Responses to Changing the Earnings Test

In this section, we examine the labor supply response to the elimination of the earnings test. The first approach examines the longitudinal response to the 2000 elimination using BEPUF data. The second approach employs a difference-in-differences framework to examine the aggregate response to the 1983 elimination for workers aged 70–71 and the 2000 elimination for workers aged 65–69.

A. Longitudinal Responses to the 2000 Elimination of the Earnings Test

With the BEPUF data, we examine how the earnings growth of 66–68-year-olds differs before and after the 2000 elimination by initial earnings levels. Figure 8 presents mean earnings growth from year T to $T + 1$ by year T bins for 66–68-year-olds. The figure shows the mean one-year growth rates for three sets of initial years: 1997–98, 1999–2000, and 2001–2002. For example, the 1997–98 period contains the mean growth rates between 1997 and 1998 and between 1998 and 1999. Theory predicts that, when the earnings test is eliminated, workers with earnings at the threshold should increase their earnings relative to workers with earnings at other points on the budget constraint and relative to workers at the threshold during other time periods. Since the earnings test was eliminated in early 2000, we look for elimination

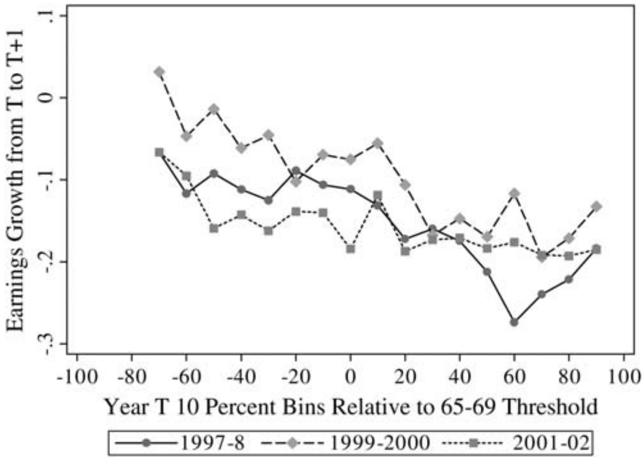

Figure 8
Average Earnings Growth from Year T to T+1 by Year T Bin, BEPUF Data

effects between 1999 and 2000 and between 2000 and 2001, under the assumption that the response to the elimination unfolded over two years.[22]

As shown in Figure 8, one-year earnings growth for 66–68-year-olds is greater in the 1999–2000 period than it is in the 1997–98 and 2001–2002 periods. Statistically significant larger earnings growth rates are found in the −70, −60, −50, −40, −30, −10, 0, 20, and 60 bins.[23] The larger earnings growth rates below the threshold are consistent with the presence of rigidities, and the larger growth rates above the threshold are consistent with the substitution effect dominating the income effect.

B. The Aggregate Labor Supply Response to the 1983 and 2000 Eliminations

To examine aggregate labor supply effects, we analyze the 1983 elimination for 70–71-year-olds and the 2000 elimination for 65–69-year-olds. Our difference-in-differences approach is similar to that used by Gruber and Orszag (2003), Tran (2003), and Song (2004).[24] Our basic regression model is

$$(1)\quad Y_{it} = \beta_0 + \beta_1 ELIM_{it} + \beta_2 Age_{it} + \beta_3 Year_t + \beta_4 X_{it} + \varepsilon_{it}$$

where $ELIM_{it}$ is an indicator variable for whether individual i in year t is in an age group for whom the earnings test was eliminated, Age_{it} is a vector of age dummies,

22. Tabulations (not shown) provide evidence that the earnings growth rates between 1999 and 2000 and between 2000 and 2001 are similar.

23. Similar to our previous statistical tests, we compute two-tailed *t*-tests bin by bin, comparing the one-year growth rates for 1999–2000 to the pooled growth rates for 1997–98 and 2001–2002. The *t*-tests indicated that the growth for 1999–2000 was greater and statistically significant at the 0.05 confidence level for the -70, -60, -50, -40, -30, -10, 0, 20, and 60 bins; the growth rate was in no case negative and statistically significant.

24. The regression model we use is a simplified version of the one used in previous studies. For example, Gruber and Orszag (1983) use variation that stems from the 1983 elimination and the incremental increases in the threshold. Nonetheless, we obtain results that are similar to those reported in the other reduced-form studies.

$Year_t$ is a vector of year dummies, and, when available, X_{it} is a vector of demographic controls including race/ethnicity (white, black, Hispanic, and other), completed education (less than high school, high school, some college, and four or more years of college), and marital status (married, divorced, widowed, and never married). We examine five outcome variables (Y_{it}): annual earnings, worked at all, log hourly wages, weeks per year, and hours per week. In the case of log hourly wages, weeks per year, and hours per week, we limit our sample to those who reported positive earnings. Earnings and wages are adjusted to 2004 dollars with the personal consumption expenditure deflator. The coefficient of interest is β_1, which measures the effect of eliminating the earnings test on the affected age group.

Our analysis uses data for the years 1998–2002 to examine the 2000 elimination with the BEPUF, 1996–2003 to examine the 2000 elimination with the CPS, and 1978–86 to examine the 1983 elimination with the CPS. The selection of a wider window of years for the CPS is motivated by sample size considerations. With respect to ages, we include 70–76-year-olds for the 1983 elimination and 66–74-year-olds for the 2000 elimination. These age ranges are selected so that our comparison group is men who are up to four years older than the affected age group. We use older men as the comparison because they are not covered by the earnings test in the period we analyze. Moreover, a life-cycle model of labor supply predicts that individuals who are younger than the ages for which the earnings test is eliminated should also respond to the change.

We show results for the aggregate effects in Table 4. The top panel shows results for the 2000 elimination with the BEPUF data. The results suggest that the elimination increased earnings by \$464 among 66–69-year-olds, or an increase of almost 8 percent (\$464/\$6,067). The second panel shows results using the CPS and suggests a much larger effect of the elimination: an increase of \$1,326 or 16 percent (\$1,326/\$8,328). The much larger effect in the CPS is at least somewhat attributable to the BEPUF earnings data being censored at the taxable limit. When we censor the CPS data in a similar fashion (the third panel), we obtain results more comparable to the BEPUF, \$778 or about a 12 percent increase. The second panel also presents evidence of whether the increase in earnings is due to an increase in working, weeks worked per year, or hours worked per week. The only significant effect is on the hours worked per week. Again, we emphasize that these aggregate estimates represent the average response across the population, which theory predicts include positive, negative, and null responses. In contrast to the 2000 elimination findings, the final panel shows that there was little systematic effect of the 1983 elimination.

Our results for aggregate effects are consistent with the findings from previous studies. Our estimates for the 2000 elimination are consistent with Tran (2003) using CPS data and Song (2004) using SSA data. Our 2000 elimination results are also consistent with the findings for eliminating an earnings test for 65-69 year olds in Canada (Baker and Benjamin 1999) and the United Kingdom (Disney and Smith 2002). Moreover, our findings of no effect for the 1983 elimination are consistent with the results of Gruber and Orszag (2003).[25]

25. Leonesio (1990) reports that Packard (1988) also found no effect of the 1983 elimination on the employment of 70–71-year-old men.

Table 4
The Effect of Eliminating the Earnings Test on Aggregate Labor Supply

			Sample restricted to working men		
	Earnings	Working	Log wages	Weeks/ year	Hours/ week
2000 Elimination for 65–69-year-olds, BEPUF 1998–2002, censored earnings data					
Dependent variable mean	6,067	0.33	—	—	—
ELIM	464**	−0.006*			
	(114)	(0.003)			
R-squared	0.02	0.02			
Sample size	300,979	300,979			
2000 Elimination for 65–69-year-olds, CPS 1996–2003, uncensored earnings data					
Dependent variable mean	8,328	0.25	2.64	42.1	31.4
ELIM	1,326**	0.01	0.07	−0.98	1.55**
	(686)	(0.01)	(0.07)	(0.73)	(0.70)
R-squared	0.04	0.04	0.06	0.01	0.02
Sample size	30,741	30,741	7,813	7,813	7,813
2000 Elimination for 65–69-year-olds, CPS 1996–2003, censored earnings data					
Dependent variable mean	6,314	—	—	—	—
ELIM	778**				
	(364)				
R-squared	0.06				
Sample size	30,741				
1983 Elimination for 70–71-year-olds, CPS 1978–86, uncensored earnings data					
Dependent variable mean	3,159	0.20	2.16	38.3	28.2
ELIM	−330	−0.01	−0.10	1.95*	0.35
	(356)	(0.01)	(0.11)	(1.13)	(1.06)
R-squared	0.04	0.02	0.05	0.02	0.02
Sample size	19,595	19,595	3,919	3,919	3,919

Notes: The dependent variable is listed at the top of each column, and the sample information is listed for each panel. In addition to the elimination/age dummy variables, each regression includes a vector of age dummy variables and a vector of year dummy variables. The CPS regressions additionally include a vector of race/ethnicity dummy variables (white, black, Hispanic, and other), a vector of completed education (less than high school, high school, some college, and 4+ years of college), and marital status (married, divorced, widowed, and never married). * and ** indicate statistical significance at the 0.10 and 0.05 level, respectively.

Table 5
The Effect of Eliminating the 2000 Earnings Test on Aggregate Earnings and Employment by Age

	BEPUF censored	CPS uncensored
Dependent variable mean	6,067	8,328
ELIM × age 66	717**	1976*
	(177)	(1076)
ELIM × age 67	489**	1711
	(179)	(1,098)
ELIM × age 68	323*	1593
	(181)	(1,112)
ELIM × age 69	305	−65
	(184)	(1,120)
R-squared	0.02	0.04
Sample size	300,979	30,741

Notes: The dependent variable is listed at the top of each column, and the sample information is listed for each panel. In addition to the elimination/age dummy variables, each regression includes a vector of age dummy variables and a vector of year dummy variables. The CPS regressions additionally include a vector of race/ethnicity dummy variables (white, black, Hispanic, and other), a vector of completed education (less than high school, high school, some college, and 4+ years of college), and marital status (married, divorced, widowed, and never married). * and ** indicate statistical significance at the 0.10 and 0.05 level, respectively.

Given the consistency in results across studies, the substantive question that arises is the following: Why do we see effects for the 2000 elimination, but not for the 1983 elimination, especially in light of the substantial evidence of bunching throughout the entire time period? A potential explanation for the different results between 1983 and 2000 is that the younger workers affected by the 2000 elimination (65–69 year olds) could more easily increase labor supply than could the 70–71 year olds affected by the 1983 elimination.[26] To examine this hypothesis, we re-estimate the earnings and employment regressions for the 2000 elimination allowing the estimated effect to vary by age. The results in Table 5 show that the effects of the 2000 elimination are monotonically decreasing in age in both data sets, although the estimated standard errors are quite large with the CPS data.[27] Thus, given the evidence of bunching throughout the time period and the evidence that only younger individuals responded to the 2000 elimination, we conclude the null result for the 1983 elimination is attributable to older ages affected by that reform.

26. Another potential explanation is that the 1983 elimination affected a population that was relatively weighted toward individuals who would have had an incentive to reduce their labor supply following elimination. However, the data suggest that the population of workers who would have had an unambiguous incentive to reduce their labor supply was large for the 2000 elimination when compared to the 1983 elimination. There are more high earners among the 65–69-year-olds in 1999 than there are among the 70–71-year-olds in 1982.

27. Tran (2003) also reports finding larger employment effects for younger workers in his CPS sample.

VII. Conclusion and Discussion

Our empirical results suggest that the Social Security earnings test has a substantial effect on male labor supply. Our bunching analyses indicate that at least 4.8 percent of workers adjust their earnings in response to the earnings test, an estimate that is about three times greater than would be obtained using previous methods. We obtain a larger estimate because our administrative data allow us to assess the extent to which measurement error and labor market rigidities obfuscate bunching at the threshold. Our longitudinal analyses indicate that men increase their earnings when they no longer face the earnings test, providing further evidence that some men were adjusting their labor supply and that labor market rigidities kept some constrained men from locating precisely at the kink. Finally, our analyses of the aggregate labor supply effects of the earnings test suggest that younger men react more to earnings test reforms than do older men. These conclusions are consistent with past studies on the earnings test in the United States and in other countries and accord with the widely held view that the earnings test imposes a substantial tax on earnings.

Our results have implications for current Social Security policies. The earnings test still covers individuals aged 62 to the full retirement age; the full retirement age is currently scheduled to increase to age 67 by 2022. Given that our results suggest individuals younger than the normal retirement age bunched at the threshold during 2000–2004 despite the actuarial fairness of how benefits were returned to workers (through the ARF), there is little reason to suspect that individuals subject to the earnings test will stop bunching at the threshold in the foreseeable future. Moreover, because our findings indicate that younger workers are most responsive, future reforms that target individuals under the full retirement age could have even larger effects than those for the 2000 elimination.

Our results also have important methodological implications. Much research is devoted to other policies that induce kinks in the budget constraint, such as the Earned Income Tax Credit (EITC) and Temporary Assistance for Needy Families (TANF). Our results suggest that the use of self-reported data can mask the degree to which labor supply responds to the incentives generated by these policies. In addition, whether one uses self-reported or administrative data, our findings also suggest that labor market rigidities can induce labor supply changes that are much more complicated than are predicted by simple models that assume frictionless labor markets. Ignoring measurement error and rigidities can result in misleading inferences regarding how policy affects labor supply.

Appendix 1
Supplementary Tables

Table A1
Descriptive Characteristics of the CPS Sample

	Full sample				SS recipients		
Year/Age	Sample size	Working	Mean Earnings	SS Receipt	Sample size	Working	Mean Earnings
1975–77							
63–64	3,344	0.574	5,714	0.611	2,040	0.390	1,685
65–69	5,317	0.342	2,123	0.861	4,588	0.295	977
70–71	2,051	0.271	1,342	0.907	1,846	0.251	931
72–76	4,166	0.226	1,183	0.922	3,841	0.226	1,107
1978–82							
63–64	6,296	0.520	7,469	0.627	3,928	0.336	2,140
65–69	10,229	0.325	3,026	0.873	8,920	0.281	1,439
70–71	4,305	0.229	1,760	0.916	3,933	0.217	1,219
72–76	8,335	0.189	1,296	0.932	7,754	0.188	1,182
1983–89							
63–64	8,600	0.456	9,606	0.665	5,729	0.293	2,674
65–69	14,518	0.290	3,971	0.884	12,861	0.256	2,118
70–71	6,071	0.204	2,318	0.938	5,707	0.203	2,181
72–76	11,665	0.161	1,774	0.942	11,015	0.162	1,676
1990–99							
63–64	10,019	0.465	13,998	0.666	6,710	0.315	4,922
65–69	18,369	0.288	6,789	0.872	16,029	0.254	3,963
70–71	8,297	0.211	4,290	0.916	7,632	0.205	3,784
72–76	16,921	0.159	3,196	0.928	15,736	0.158	2,900
2000–2004							
63–64	6,320	0.486	20,418	0.643	4,044	0.339	9,018
65–69	10,150	0.319	11,695	0.878	8,835	0.302	10,411
70–71	4,634	0.254	7,995	0.905	4,179	0.241	7,069
72–76	10,268	0.171	5,093	0.912	9,302	0.164	4,448

Table A2
Descriptive Characteristics of the NBDS Retired Workers Sample, 1983–89

	Full sample		Current beneficiary sample	
	N	Working	N	Working
63	497	0.254	496	0.252
64	1,471	0.240	1,466	0.239
65	1,432	0.209	1,421	0.206
66	1,900	0.216	1,888	0.214
67	2,764	0.209	2,748	0.208
68	3,991	0.223	3,961	0.221
69	4,280	0.219	4,251	0.217
70	4,322	0.209	4,316	0.209
71	3,779	0.192	3,776	0.192
72	2,862	0.210	2,860	0.210
73	2,361	0.200	2,360	0.200
74	1,535	0.206	1,534	0.206
75	461	0.182	460	0.183
76	133	0.198	132	0.200

Table A3
Descriptive Characteristics of the BEPUF Sample, 1990–2003

	Full sample		Beneficiary sample	
Age	Sample size	Fraction working	Sample size	Fraction working
63	104,199	0.568	42,562	0.377
64	103,206	0.521	53,602	0.378
65	100,050	0.486	57,728	0.366
66	95,464	0.431	89,981	0.409
67	91,069	0.390	87,549	0.375
68	86,643	0.360	84,153	0.349
69	82,068	0.333	80,267	0.325
70	77,354	0.308	76,109	0.303
71	72,627	0.285	72,090	0.283
72	67,598	0.262	67,206	0.261
73	62,582	0.243	62,274	0.242
74	57,099	0.222	56,877	0.222
75	51,954	0.202	51,783	0.202
76	46,751	0.185	46,608	0.184

Table A4
The Amount and Statistical Significance of Bunching in the Figures

	Panel of Figure					
	A	B	C	D	E	F
Figure 2						
66–69 workers	0.86	0.81	0.72	0.72	0.58	
	[0.81, 0.90]	[0.77, 0.86]	[0.67, 0.76]	[0.67, 0.76]	[0.51, 0.65]	
	216	333	347	394	188	
70–71 workers	0.83	0.80	0.58	0.67	0.57	
	[0.71, 0.92]	[0.72, 0.87]	[0.47, 0.69]	[0.56, 0.77]	[0.43, 0.69]	
	59	111	81	91	60	
72–76 workers	0.67	0.67	0.56	0.55	0.51	
	[0.54, 0.78]	[0.58, 0.76]	[0.45, 0.67]	[0.46, 0.63]	[0.40, 0.63]	
	63	110	82	140	78	
Figure 3						
63–64 workers	0.80	0.66	0.71	0.65	0.64	
	[0.72, 0.87]	[0.58, 0.73]	[0.65, 0.78]	[0.59, 0.71]	[0.55, 0.71]	
	121	152	203	235	149	
66–69 workers	0.86	0.55	0.55	0.55	0.50	
	[0.81, 0.90]	[0.49, 0.61]	[0.49, 0.60]	[0.50, 0.61]	[0.43,0.58]	
	216	272	315	316	177	

Figure 4						
63–64 workers	0.73	0.68	0.67	0.85	0.83	0.79
	[0.67, 0.80]	[0.61, 0.75]	[0.57, 0.76]	[0.76, 0.92]	[0.82, 0.85]	[0.77, 0.81]
	177	186	98	86	3,044	1,343
71–74 workers	0.57	0.52	0.59	0.53	0.54	0.54
	[0.48, 0.66]	[0.43, 0.60]	[0.48, 0.69]	[0.45, 0.61]	[0.52, 0.56]	[0.51, 0.56]
	130	128	85	165	2,462	1,632
Figure 5						
66–69 workers	0.74	0.75	0.85	0.79		
	[0.69, 0.79]	[0.69, 0.79]	[0.81, 0.88]	[0.78, 0.80]		
	324	350	398	6,249		
71–74 workers	0.60	0.58	0.57	0.59		
	[0.49, 0.70]	[0.49, 0.67]	[0.49, 0.65]	[0.57, 0.62]		
	95	131	175	2,086		

Notes: As a simple measure of bunching, this table calculates the fraction of workers in the -10 bin compared to the total workers in the -10 and 10 bins. For each graph point, the first line is the point estimate of the fraction, the second line (in brackets) is the 95 percent confidence interval based on a binomial distribution, and the third line is the total number of workers in both bins.

References

Baker, Michael, and Dwayne Benjamin. 1999. "How Do Retirement Tests Affect the Labour Supply of Older Men?" *Journal of Public Economics* 71(1):27–52.

Blinder, Alan S., Gordon, Roger H., and Donald E. Wise. 1980. "Reconsidering the Work Disincentive Effects of Social Security." *National Tax Journal* 33(4):431–42.

Blundell, Richard, and Thomas MaCurdy. 1999. "Labor Supply: A Review of Alternative Approaches." In *Handbook of Labor Economics* (Volume 3A), eds. Orley Ashenfelter and David Card, 1559–695. Amsterdam: Elsevier Science.

Borjas, George. 2005. *Labor Economics* (Third Edition). Boston: McGraw-Hill Irwin Press.

Bound, John, and Alan B. Krueger. 1991. "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics* 9(11):1–24.

Burkhauser, Richard V., and John A. Turner. 1978. "A Time Series Analysis on Social Security and Its Effects on the Market Work of Men at Younger Ages." *Journal of Political Economy* 86(4):701–16.

Burtless, Gary, and Robert A. Moffitt. 1985. "The Joint Choice of Retirement Age and Post-Retirement Hours of Work." *Journal of Labor Economics* 3(2):209–36.

Disney, Richard, and Sarah Smith. 2002. "The Labor Supply Effect of the Abolition of the Earnings Rule for Older Workers in the United Kingdom." *Economic Journal* 112(478):C136–52.

Friedberg, Leora. 2000. "The Labor Supply Effects of the Social Security Earnings Test." *The Review of Economics and Statistics* 82(1):46–63.

Gruber, Jonathan, and Peter Orszag. 2003. "Does the Social Security Earnings Test Affect Labor Supply and Benefits Receipt?" *National Tax Journal* 56(4):755–73.

Gustman, Alan, and Thomas Steinmeier. 1983. "Minimum Hours Constraints and Retirement Behavior." *Contemporary Policy Issues* 3(April):77–91.

———. 1985. "The 1983 Social Security Reforms and Labor Supply Adjustments of Older Individuals in the Long Run." *Journal of Labor Economics* 3(2): 237–53.

———. 1991. "Changing the Social Security Rules for Work after 65." *Industrial and Labor Relations Review* 44(4):733–45.

———. 2004. "The Social Security Retirement Earnings Test, Retirement and Benefit Claiming." NBER Working Paper 10905.

Haider, Steven J., and David Loughran. 2007. "Do Elderly Men Respond to Taxes on Earnings? Evidence from the Social Security Retirement Earnings Test." RAND Corporation Working Paper WR-223–1.

Honig, Marjorie, and Cordelia Reimers. 1989. "Is it Worth Eliminating the Retirement Earnings Test?" *American Economic Review Papers and Proceedings* 79(2):103–107.

Hurd, Michael. 1996. "The Effect of Labor Market Rigidities on the Labor Force Behavior of Older Workers." In *Advances in the Economics of Aging*, ed. David Wise, 11–58. Chicago: University of Chicago Press.

Kaufman, Bruce, and Julie Hotchkiss. 2006. *The Economics of Labor Markets* (Seventh Edition). Mason, Oh.: Thomson South-Western.

Leonesio, Michael V. 1990. "The Effects of the Social Security Earnings Test on the Labor-Market Activity of Older Americans: A Review of the Empirical Evidence." *Social Security Bulletin* 53(5):2–21.

Lundberg, Shelly. 1985. "Tied Wage-Hours Offers and the Endogeneity of Wages." *The Review of Economics and Statistics* 67(3):405–10.

Packard, Michael D. 1988. "The Effects of Removing 70- and 71-Year-Olds from Coverage Under the Social Security Earnings Test." Office of Policy, Social Security Administration. Unpublished.

Pingle, Jonathan. 2006. "Social Security's Delayed Retirement Credit and the Labor Supply of Older Workers." Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series 2006–37.

Reimers, Cordelia, and Marjorie Honig. 1993. "The Perceived Budget Constraint under Social Security: Evidence from Reentry Behavior." *Journal of Labor Economics* 11(1):184–204.

Rust, John, and Christopher Phelan. 1997. "How Social Security and Medicare Affect Retirement Behavior in a World of Incomplete Markets." *Econometrica* 65(4):781–831.

Saez, Emmanuel. 2002. "Do Taxpayers Bunch at Kink Points?" Department of Economics, University of California Berkeley. Unpublished.

Social Security Administration. 2004. *Annual Statistical Supplement*. Washington, D.C.: Office of Policy, Social Security Administration.

Song, Jae G. 2004. "Evaluating the Initial Impact of Eliminating the Retirement Earnings Test." *Social Security Bulletin* 65(1):1–15.

Tran, Bac V. 2003. "The Effect of the Repeal of the Retirement Earnings Test on the Labor Supply of Older Workers." University of Maryland. Unpublished.

Viscusi, Kip W. 1979. *Welfare of the Elderly: An Economic Analysis and Policy Prescription*. New York: John Wiley & Sons.

Vroman, Wayne. 1985. "Some Economic Effects of the Retirement Test." In *Research in Labor Economics* (Volume 7), ed. Ronald Ehrenberg 31–89. Greenwich, Conn.: JAI Press Inc.