Disability Risk, Disability Insurance and Life Cycle Behavior*

Hamish Low
University of Cambridge and IFS

Luigi Pistaferri
Stanford University, NBER, CEPR and IZA

December 10, 2009

Abstract

This paper provides a life-cycle framework for weighing up the insurance value of disability benefits agains the incentive cost. Within this framework, we estimate the life-cycle risks that individuals face in the US, as well as the parameters governing the disability insurance program, using indirect inference and longitudinal data on consumption, disability status, disability insurance receipt, and wages. We use our estimates to characterize the effectiveness of the disability insurance program and to consider what policy reforms increase welfare. The high Type I errors associated with the screening process lead to increases in welfare as the program becomes less strict, despite the increase in moral hazard that this implies.

1 Introduction

The Disability Insurance (DI) program in the US is a large and rapidly growing social insurance program offering income replacement and health care benefits to people with working disabilities. In 2007, the cash benefits paid by the DI program were three times larger than those paid by Unemployment Insurance (UI) ($99.1 billions vs. $32.5 billions).[1] Between 1985 and 2007 the proportion of DI claimants in the US has almost doubled (from about 2.5% to almost 5% of the working-age population, see Autor and Duggan, 2006). The key questions in thinking about the size and growth of the program are whether program claimants are genuinely unable to work, and how valuable is the insurance provided. These questions underlie the concern that the greater use of DI explains the recent decline in labor market participation of men and the concern that the DI program is being used as a gateway for early retirement, rather than providing insurance against health shocks that prevent work. To evaluate these concerns and to evaluate the costs and

*Low thanks funding from the ESRC under an ESRC Research Fellowship. Pistaferri thanks funding from NIH/NIA under grant 1R01AG032029-01 and from NSF under grant SES-0921689. We have received useful comments from audiences at various conferences and departments in Europe and the US. We are especially grateful to Costas Meghir and Aleh Tsyvinski for detailed comments, and to Katja Kaufmann and Itay Saporta for research assistance. All errors are our own.

[1] The relative size of DI is even larger if we add the in-kind health care benefits provided by the Medicare program.

benefits of changing the DI program to try to reduce disincentives to work, we need a realistic framework of the risks that individuals face over their life-cycle to their health and ability to work, and a framework that captures both the insurance benefit of DI as well as the incentive effects on individual behavior. The broad aim of this paper is to provide this quantitative evaluation of the DI program in an explicit life-cycle setting.

More specifically, our paper has three goals. First, we propose a theoretical framework that allows us to study in a life cycle setting the effect of disability risk on labor supply, savings and the DI application decisions jointly. We consider the problem of an individual who faces two types of shock to wages: a permanent productivity shock unrelated to health; and a "disability" shock which reduces the return to work. The distinction between the two types of shock to wages is key for understanding the incentive problem with the DI program:[2] an individual with a disability shock above a certain threshold can not work; while an individual with a productivity shock below a certain threshold may not want to work. Either may apply for DI benefits. Whether such labor supply distortions occur depends on a number of factors, such as the extent of labor market frictions and the availability of alternative forms of insurance (own savings, as well as other government-provided insurance programs).

The second goal is to estimate the parameters of this model using microeconomic data. We use PSID data on wages and indicators of disability status to help identify the parameters of the wage process, data on consumption changes upon disability to identify preference parameters, and data on employment by age and disability status to identify labor market frictions. We identify the policy parameters governing the disability application and review process using data on the stock of individuals on DI (by age and work limitation status) and data on the DI status of people of different age and health status. Our estimates highlight that there are substantial false negatives in the allocation of disability insurance, while false positives are somewhat less problematic.

Finally, we use our model and the estimates of the structural parameters to tackle the welfare and policy questions directly. The model can be used to address a number of questions, such as how well insured are individuals against disability risk, or how responsive are labour supply and savings to changes in the details of the DI program. The ability to evaluate these questions in

[2] To be consistent with the terminology adopted in the literature (e.g., Bound and Burkhauser, 1999), we label the incentive effect "moral hazard" even though there is no hidden action.

a coherent unified framework is one of the main benefits of the paper. We conduct a number of counterfactual experiments, associated to changes in: (a) stringency of the screening process, (b) re-assessment rate, (c) replacement rate, (d) generosity of alternative social insurance programs. The most striking finding of our paper is that the high Type I errors associated with the screening process of the disability insurance program lead to ex-ante welfare rising when the program becomes less strict, despite the increase in moral hazard that this implies.

Some of the issues raised in this paper have been addressed elsewhere in the literature. Some authors determine the work disincentive effect of DI by computing how many DI recipients would be in the labor force in the absence of the program. Gastwirth (1972) computes the labor force participation rate of people with disabilities who are not on DI (86%), and concludes that it is an upper bound for the proportion of DI recipients who would be working in the absence of the DI program. It is not clear, however, whether this is the best "control group" for DI beneficiaries. Parsons (1980) shows that higher DI replacement rates are associated with a large fall in male labour force participation. By contrast, Bound (1989) assumes that the correct comparison group is DI applicants who were rejected. He finds that only 1/3 to 1/2 of rejected applicants are working during his sample period, and given that this group is composed of individuals who are presumably healthier on average than DI beneficiaries, their working behavior is an upward bound for how many DI beneficiaries would be working in the absence of the program. These lower estimates have recently been confirmed by Chen and van der Klaauw (2008). Kreider (1999) uses a more structural approach to understand the joint decision of applying for DI and of not participating in the labour market, and finds that although DI has important disincentive effects on labour supply, the change in DI generosity cannot fully explain the fall in labour force participation.

More recently, Autor and Duggan (2003) provide a detailed analysis of the trends in DI receipt and the causes behind these trends and conclude that the growth is due to an (unintended) increase in DI benefit generosity (especially for people at the bottom of the wage distribution) and more lax screening. They also calculate that the U.S. unemployment rate would be two-thirds of a percentage point higher were it not for the liberalized disability system.

Given that the true disability status of an individual is private information, the presence of moral hazard effects found in the studies cited above implies that DI evaluators are prone to make two types of errors: awarding benefits to undeserving applicants, or denying them to truly disabled individuals. How large are these errors? An earlier attempt at measuring such errors is Nagi (1969), who uses a sample of 2,454 initial disability determinations. The individuals in his sample were intensely examined by a team of doctors, psychologists, social workers, etc. Nagi (1969) concluded that about 19% of those initially awarded benefits were undeserving, and 48% of those denied were truly disabled. These numbers are roughly consistent with those in Benitez-Silva et al. (2007), who use HRS data on DI application and appeal process, DI award, and self-reported information on disability status. Their conclusion is that over 40% of recipients of DI are not truly work limited and this adds to the picture of an inefficient insurance program.

The broader issue of the value of DI requires an evaluation of the benefit of the insurance provided by DI as well as an assessment of the efficiency loss. Hoynes and Moffitt (1997) examine work disincentives for DI beneficiaries, and conclude via simulations that some of the reforms aimed at allowing DI beneficiaries to keep more of their earnings are unlikely to be successful and may, if anything, increase the number of people applying for DI. In a similar vein, Acemoglu and Angrist (1998) and DeLeire (2000) examine the effect of the American with Disabilities Act, which should have eased the transition back to work of the disabled, and find that it actually led to a decline in the employment rate of people with disabilities, perhaps because of the burden imposed onto employers. Previous work by Bound et al. (2004), Waidmann et al. (2003) and Rust et al. (2002) has also highlighted the importance of considering both sides of the insurance/incentive trade-off for welfare analysis. Our work builds on these papers but improves on their framework by having a more flexible specification of preferences and the wage process, and by adding labor market frictions and interactions with other social insurance programs. None of these elements are purely cosmetic. Allowing for non-separability between consumption and bad health may explain a fall of consumption upon disability even in the presence of full insurance. Non-health related negative productivity shocks and lack of employment opportunities are at the root of the moral hazard problem: both reduce the opportunity cost of applying for DI conditional on health. Finally, the opportunity cost of applying for DI depends on whether there are programs to finance consumption during the period it takes for an application to be processed. A full model of behavior is necessary to evaluate the effectiveness and value of proposed reforms.[3]

One such proposal is the Golosov and Tsyvinski (2006) proposal to impose an asset test on disability applicants. Golosov and Tsyvinski show that, if disability status is private information, an asset test can implement the constrained Pareto optimum. This result may depend on the assumption that assets are used to smooth periods of non-employment associated with disability. In our framework, where assets held for precautionary reasons are substitutable with assets held for life-cycle reasons, the welfare benefit of an asset test is ambiguous. Alternative proposals include increasing the medical hurdle for applicants, raising the reassessment rate among recipients, increasing the waiting time before a DI application is permitted, lowering the cost of work and increasing the availability of job opportunities for DI beneficiaries. The analysis of policy in a life-cycle framework as in Hubbard et al. (1995) has not, however, explicitly considered health risks, which may differ in important ways from productivity risk, or modelled the disability insurance program.

There have been some recent papers identifying the extent of health risk. In particular, DeNardi, French and Jones (2006) estimate the risk to health expenditure, but their focus is on the elderly, rather than those of working age when disability insurance is an active option. Adda, Banks and Gaudecker (2006) estimate the effect of income shocks on health and find only small effects. Recently, Meyer and Mok (2007) and Stephens (2001) have estimated in a reduced form way the effect of disability on household consumption. The value of our paper is in combining estimates of the risk associated with health shocks in a framework that allows the evaluation of the social insurance provided by DI. Gallipoli and Turner (2009) do a more refined exercise.

The rest of the paper is as follows. Section 2 presents the life-cycle model allowing for health status, and discusses the various social insurance programs available to individuals. Section 3 discusses the identification strategy. Section 4 summarizes the data used in the estimation of the model, focusing on the data on disability status and on consumption. Section 5 presents the estimates of the structural parameters. Section 6 discusses the implications of the results for the optimality of the parameters of the disability insurance program and section 7 concludes.

[3] See also Diamond and Sheshinski (1995) for a model of optimal disability insurance.

2 Life-Cycle Model

2.1 Individual Problem

We consider the problem of an individual who maximizes lifetime expected utility:

$$\max_{c,P,DI^{App}} V_{it} = E_t \sum_{s=t}^{T} \beta^{s-t} U(c_{is}, P_{is}; L_{is})$$

where β is the discount factor, E_t the expectations operator conditional on information available in period t (a period being a quarter of a year), P a discrete $\{0,1\}$ labor supply participation variable, c_t consumption, and L_t a discrete disability status indicator $\{0,1,2\}$.[4] Individuals live for T periods, may work T^W years (from age 23 to 62), and face an exogenous mandatory spell of retirement of T^R =10 years at the end of life. The date of death is known with certainty and there is no bequest motive.

The intertemporal budget constraint during the working life has the form

$$A_{it+1} = R \left[\begin{array}{c} A_{it} + \left(w_{it} h \left(1-\tau_w\right) - F\left(L_{it}\right)\right) P_{it} \\ +(B_{it} E_{it}^{UI} \left(1 - E_{it}^{DI}\right) + DI_{it} E_{it}^{DI} + S_{it} E_{it}^{DI} E_{it}^{W}) \left(1-P_{it}\right) \\ +W_{it} E_{it}^{W} - c_{it} \end{array} \right]$$

where A are beginning of period assets, R is the interest factor, w the hourly wage rate, h a fixed number of hours (corresponding to 500 hours per quarter), τ_w a proportional tax rate that is used to finance social insurance programs, F the fixed cost of work that depends on disability status,[5] B unemployment benefits, W the monetary value of the means tested welfare payment, DI the amount of disability insurance payments obtained, S_{it} the amount of SSI benefits, and E^{UI}, E^{DI}, and E^W are recipiency $\{0,1\}$ indicators for unemployment insurance, disability insurance, and the means-tested welfare program, respectively.

The worker's problem is to decide whether to work or not. When unemployed he has to decide whether to accept a job that may have been offered or wait longer. If eligible, the unemployed person will have the option to apply for disability insurance. Whether employed or not, the individual has to decide how much to save and consume. Accumulated savings can be used to finance spells out of work and retirement.

[4] This is dictated by the data we have (see below).

[5] The fact that disabled individuals face direct costs of work is explicitly recognized by the SSA, which allows individual to deduct costs of work (such as "a seeing eye dog, prescription drugs, transportation to and from work, a personal attendant or job coach, a wheelchair or any specialized work equipment") from monthly earnings before determining eligibility for DI benefits (see SSA Publication No. 05-10095).

We use a utility function of the form

$$u\left(c_{it}, P_{it}; L_{it}\right) = \frac{\left(c_{it} \exp\left(\theta L_{it}\right) \exp\left(\eta P_{it}\right)\right)^{1-\gamma}}{1-\gamma}$$

We impose that $\gamma > 1$ (in particular, we set $\gamma = 1.5$ following Attanasio and Weber, 1995), and estimate θ and η. To be consistent with disability and work being "bads", we require $\theta < 0$ and $\eta < 0$, two restrictions that are not rejected by the data. The parameter θ captures the utility loss for the disabled in terms of consumption. Participation also induces a utility loss determined by the value of η. This implies that consumption and participation are Frisch complements (i.e. the marginal utility of consumption is higher when participating) and that the marginal utility of consumption is higher when suffering from a work limitation.[6]

We assume that individuals are unable to borrow: $A_{it} \geq 0$. In practice, this constraint has bite because it precludes borrowing against unemployment insurance, against disability insurance, against social security and against the means-tested program.

At retirement, people collect social security benefits which are paid according to a formula similar to the one we observe in reality (see below). These benefits, along with assets that people have voluntarily accumulated over their working years, are used to finance consumption during retirement.

There are important differences by skills both in terms of probability of disability shocks and disability insurance recipiency rates (see the Web Appendix for more details).[7] In particular, if we proxy skills with education, we find that individuals with low education (at most high school degree) and high education (some college or more), have very similar DI recipiency rates until their mid 30s, but then there is a wide difference opening up. By age 60, the low educated are four times more likely to be DI claimants than the high educated (16% vs. 4%). In part, this is due to the fact that low educated individuals are more likely to have a severe disability at all ages. To account for these differences, in what follows we assume that all the parameters of the model are education-specific. To simplify notation, we omit subscripts defining the skill group of interest.

[6] See Finkelstein et al. (2008) for a recent attempt to measure the effect of health status on the marginal utility of consumption.

[7] The Web Appendix is available at http://www.stanford.edu/~pista/papers/WA_LP.pdf.

2.2 The Wage Process and Labour Market Frictions

We model the wage process for individual i as being subject to general productivity shocks and shocks to the disability status (as well as the contribution of observable characteristics X_{it} - which we take to be a quadratic in age):

$$\ln w_{it} = X_{it}'\alpha + \beta_1 L_{it}^1 + \beta_2 L_{it}^2 + \varepsilon_{it} \tag{1}$$

where $L_{it}^j = \mathbf{1}\{L_{it} = j\}$, and

$$\varepsilon_{it} = \varepsilon_{it-1} + \zeta_{it}$$

Individuals work limitation status, L_{it}, evolves according to a three state first-order Markov process which is age dependent. Upon entry into the labor market, all individuals are assumed to be healthy ($L_{i0} = 0$). Transition probabilities from any state depend on age. We assume that these transition probabilities are exogenous and in particular, we rule out the possibility of individuals investing in health prevention activities.[8]

Equation (1) determines the evolution of individual productivity. Productivity determines the offered wage when individuals receive a job offer. In our framework, individuals make a choice about whether or not to accept an offered wage. This will also depend on the fixed costs of work, which in turn depend on the extent of the work limitation, $F(L)$. In addition, there are labour market frictions which mean that not all individuals receive job offers. First, there is job destruction, δ, which forces individuals into unemployment for (at least) one period. Second, job offers for the unemployed arrive at a rate λ and so individuals may remain unemployed even if they are willing to work.

This wage and employment environment implies a number of sources of risk, from individual productivity, work limitation shocks and from market frictions. These risks are idiosyncratic, but we assume that there are no markets to provide insurance against these risks. Instead, there is partial insurance coming from government insurance programs (as detailed in the next section) and from individuals' own saving (and labor supply).

[8] We allow the process to differ by education, which may implicitly capture differences in health investments.

2.3 Social Insurance

2.3.1 The SSDI Program

The Social Security Disability Insurance program (SSDI) is an insurance program for covered workers, their spouses, and dependents that pays benefits related to average past earnings. The purpose of the program is to provide insurance against health shocks that impair substantially the ability to work. The difficulty with providing this insurance is that health status and the impact of health on the ability to work is imperfectly observed.[9]

The program was enacted in 1956 for individuals older than 50 and suffering from an impairment that was "expected to result in death or be of long, continued, and indefinite duration". In later years eligibility was extended to individuals under age 50, disability did not have to be permanent any more, waiting periods were reduced (from 6 to 5 months) and benefit levels increased. By the mid-1970s typical after-tax replacement rates reached 60%. The Social Security Administration (SSA) responded to a substantial growth in the DI roll by refining their regulations guiding decisions and by changing the frequency and nature of medical eligibility reviews for DI beneficiaries, which led to a fall in award rates from 48.8% to 33.3% between 1975 and 1980 and to an increase in the number of terminations. In 1984, eligibility criteria were liberalized, when the SSA issued new rulings that gave controlling weight to source evidence (e.g., own physician). In 1999, a number of work incentive programs for DI beneficiaries were introduced (such as the Ticket to Work program) in an attempt to push some of the DI recipients back to work.

The award of disability insurance depends on the following conditions: (1) An individual has to have filed an application for disabled worker's benefits; (2) There is a work requirement on the number of quarters of prior participation: Workers over the age of 31 are disability-insured if they have 20 quarters of coverage during the previous 40 quarters; (3) There is a statutory five-month waiting period out of the labour force from the onset of disability before an application will be processed; and (4) Finally, the individual must meet a medical requirement, i.e. the presence of a disability defined as "*Inability to engage in any substantial gainful activity by reason of any medically determinable physical or mental impairment, which can be expected to result in death, or*

[9] Besides SSDI, about 25% of workers in the private sector are also covered by employer-sponsored long-term disability insurance plans.

which has lasted, or can be expected to last, for a continuous period of at least 12 months."

This requires that the disability affects the ability to work; and further, both the severity and the expected persistence of the disability matter. The actual DI determination process consists of sequential steps. After excluding individuals earning more than a so-called "substantial gainful amount" (SGA, currently $940 for non-blind individuals) in a first step, the SSA determine whether the individual has a medical disability that is severe and persistent (per the definition above). If such disability is a listed impairment, the individual is awarded benefits without further review.[10] If the applicant's disability does not "meet" or "equal" a listed impairment, the DI evaluators try to determine the applicant's residual functional capacity. In the last stage the "pathological" criterion is paired with an "economic" criterion. Two individuals with the identical working disability may receive different DI determination decisions depending on their age, education, general skills, and even economic conditions faced at the time the determination is made.

In our model, we make the following assumptions in order to capture the complexities of the disability insurance program detailed above. First, we require that the individuals make the choice to apply for benefits. Second, individuals have to have been at work for at least one period prior to becoming unemployed and making the application. Third, individuals must have been unemployed for at least one quarter before applying. Successful applicants begin receiving benefits in that second quarter. Unsuccessful individuals must wait a further quarter before being able to return to work, but there is no direct monetary cost of applying for DI. Finally, we assume that the probability of success depends on the true work limitation status, age, and education:

$$\Pr\left(DI_{it}=1|\,DI_{it}^{App}=1,L_{it},t\right)=\begin{cases}\pi_L^{Young} & \text{if } t<45\\ \pi_L^{Old} & \text{if } 62\geq t\geq 45\end{cases}$$

The medical requirement in the SSDI program imposes a severity and persistence requirement on the work limitation. In our model, the expected persistence of the work limitation is captured by the Markov process for wages and is age dependent. This age dependence of the persistence in our model is the reason why we make the probability of a successful application for DI dependent on age.[11] The survey question survey we use (described below) asks individuals about *work*-related

[10] The listed impairments are described in a blue-book published and updated periodically by the SSA ("Disability Evaluation under Social Security"). The listed impairments are physical and mental conditions for which specific disability approval criteria has been set forth or listed (for example, "Amputation of both hands", "Heart transplant", or "Mental retardation", defined as full scale IQ of 59 or less, among other things).

[11] The separation at age 45 takes also into account the practical rule followed by DI evaluators in the the last stage

limitations rather than medical conditions or health status more generally. Finally, we account for the "economic" role played by the last step of the DI determination process by adding labor market frictions to our model and assuming that the parameters are skill-specific.

Individuals leave the disability program either voluntarily (which in practice means into employment) or following a reassessment of the work limitation and being found to be able to work. The probability of being reassessed is 0 for the first year, then is given by P^{Re}, which is independent of L and age. If an individual is not successful on application or if an individual is rejected on reassessment, the individual has to remain unemployed until the next quarter before taking up a job. Individuals can only re-apply in a subsequent unemployment spell.

SSDI benefits are calculated in essentially the same fashion as Social Security retirement benefits, and have been subject to the same changes in benefit levels. Beneficiaries receive indexed monthly payments equal to their Primary Insurance Amount (PIA), which is based on taxable earnings averaged over the number of years worked (known as AIME). Caps on the amount that individuals pay into the DI system as well as the nature of the formula determining benefits (see (2) below) make the system progressive. Because of the progressivity of the benefits and because of the fact that individuals receiving SSDI also receive Medicare benefits after two years, the replacement rates (i.e., the percentage of before-disability income an individual will receive once she ceases working) are substantially higher for workers with low earnings and those without employer-provided health insurance. However, benefits are independent of the extent of the work limitation.

In the model, we set the value of the benefits according to the actual schedule in the US program. The value of disability insurance is given by

$$D_{it} = \begin{cases} 0.9 \times \overline{w}_i & \text{if } \overline{w}_i \leq a_1 \\ 0.9 \times a_1 + 0.32 \times (\overline{w}_i - a_1) & \text{if } a_1 < \overline{w}_i \leq a_2 \\ 0.9 \times a_1 + 0.32 \times (a_2 - a_1) + 0.15 \times (\overline{w}_i - a_2) & \text{if } a_2 < \overline{w}_i \leq a_3 \\ 0.9 \times a_1 + 0.32 \times (a_2 - a_1) + 0.15\,(a_3 - a_2) & \text{if } \overline{w}_i > a_3 \end{cases} \tag{2}$$

where $\overline{w}_i$ is average earnings computed before the time of the application and a_1, a_2, and a_3 are thresholds we take from the legislation.[12] We assume $\overline{w}_i$ can be approximated by the value of the

of the DI determination process (the so-called Vocational Grid, see Appendix 2 to Subpart P of Part 404—Medical-Vocational Guidelines, as summarized in Chen and van der Klaauw, 2008).

[12] In reality what is capped is $\overline{w}_i$ (the AIME), because annual earnings above a certain threshold are not subject to payroll taxation. We translate a cap on AIME into a cap on DI payments.

permanent wage at the time of the application. Whether an individual is eligible (i.e., $E_{it}^{DI} = 1$) depends on the decision to apply ($DI_{it} = 1$) while being out of work. We assume that the probability of success is independent of age. Eligibility does not depend on whether an individual quits or the job is destroyed.

In retirement, all individuals receive social security calculated using the same formula (2) used for disability insurance.

2.3.2 Unemployment Insurance

We assume that unemployment benefits are paid only for the quarter immediately following job destruction. We define eligibility for unemployment insurance E_{it}^{UI} to mirror current legislation: benefits are paid only to people who have worked in the previous period, and only to those who had their job destroyed (job quitters are therefore ineligible for UI payments, and we assume this can be perfectly monitored).[13] We assume $B_{it} = b \times w_{it-1}\overline{h}$, subject to a cap, and we set the replacement ratio $b = 75\%$. This replacement ratio is set at this high value because the payment that is made is intended to be of a similar magnitude to the maximum available to someone becoming unemployed.

In the US, unemployment benefit provides insurance against job loss and insurance against not finding a new job. However, under current legislation benefits are only provided up to 26 weeks (corresponding to two periods of our model) and so insurance against not finding a new job is limited. Our assumption is that there is no insurance against the possibility of not receiving a job offer after job loss. This simplifying assumption means that, since the period of choice is one quarter, unemployment benefit is like a lump-sum payment to those who exogenously lose their job and so does not distort the choice about whether or not to accept a new job offer. The only distortion is introduced by the tax on wages.

[13] We have simplified considerably the actual eligibility rules observed in the US. A majority of states have eligibility rules which are tougher than the rule we impose, both in terms of the number of quarters necessary to be eligible for any UI and in terms of the number of quarters of work necessary to be eligible for the maximum duration (Meyer, 2002). However, making eligibility more stringent in our model is numerically difficult because the history of employment would become a state variable (the same is true for DI eligibility). Our assumption on eligibility shows UI in its most generous light.

2.3.3 Universal Means-Tested Program

In modelling the universal means-tested program, our intention was to mirror partially the actual food stamps program but with three important differences. First, the means-testing is only on household income rather than on income and assets; second, the program provides a cash benefit rather than a benefit in kind; and third, we assume there is 100% take-up.[14] These assumptions mean the program plays the role of providing a floor to income for all individuals. This is similar to Hubbard, Skinner and Zeldes (1995). Gross income is given by

$$y_{it}^{gross} = w_{it} h P_{it} + \left(B_{it} E_{it}^{UI} \left(1 - E_{it}^{DI}\right) + D_{it} E_{it}^{DI} \right) \left(1 - P_{it}\right) \tag{3}$$

giving net income as $y = (1 - \tau_w)\, y^{gross} - d$, where d is the standard deduction that people are entitled to when computing net income for the purpose of determining food stamp allowances. The value of the program is then given by

$$T_{it} = \begin{cases} \overline{T} - 0.3 \times y_{it} & \text{if } E_{it}^{W} = 1 \ \left(\text{i.e., if } y_{it} \leq \underline{y}\right) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The maximum value of the payment, $\overline{T}$, is set assuming a household with two adults and two children, although in our model there is only one earner. The term $\underline{y}$ should be interpreted as a poverty line. In the actual food stamp program, only people with net earnings below the poverty line are eligible for benefits ($E_{it}^{W} = 1$).

The distinction with the actual food stamps program is that the means-tested program in this paper is not asset tested. The program interacts in complex ways with disability insurance: the Food Stamps program provides a consumption floor during application for DI.

2.3.4 Supplemental Security Income (SSI)

Individuals who are deemed to be disabled according to the rules of the DI program and who have income (comprehensive of DI benefits but excluding the value of food stamps) below the threshold that would make them eligible for food stamps receive also supplemental security income (SSI). The SSI program in the US is designed to help aged and disabled people who have little or no

[14] The difficulty with allowing for an asset test in our model is that there is only one sort of asset which individuals use for retirement saving as well as for short-term smoothing. In reality, the asset test applies only to liquid wealth and thus excludes pension wealth (as well as real estate wealth and other durables).

income.[15] The definition of disability in the SSI program is identical to the one for the DI program (an individual is considered disabled if he or she has a medically determinable physical or mental impairment which results in the inability to do any substantial gainful activity, can be expected to result in death, or has lasted or can be expected to last for a continuous period of not less than 12 months). The definition of low income is similar to the one used for the Food Stamps program.[16]

2.3.5 Taxation on Earnings

The tax rate on earnings, τ_w, is set to hold the government budget in balance when varying the parameters of the social insurance policies.

2.4 Model Discussion

To understand our characterization of the application process and the trade-off between genuine applicants and non-genuine applicants, consider the following example. Assume that the government receives a noisy signal S_{it} about the true disability status of a DI applicant (independent of non-health related productivity ε_{it}), and that its decision rule is to award benefits to applicants whose signal exceed a certain stringency threshold, $S_{it} > \overline{S}$. Below this level, some individuals may wish to apply for DI if their productivity is sufficiently low because the government only observes a noisy measure of the true disability status. Symmetrically, some individuals with true disability status above the threshold may not apply because they are highly productive (they have high realizations of ε_{it}) despite their disability.

A lower level of work limitations means income is not affected significantly by them and that there is less chance of being wrongly assessed as needing DI and this implies that applications for DI will be made only by individuals with a very low level of productivity. This is the heart of the moral hazard problem.[17] Black et al. (2002) make a similar point when looking at the experience of the coal industry in the US. The industry went through a transitory boom and a permanent bust related to the oil shocks of the 1970s. Black et al. (2002) show that consistent with a model

[15] Given our focus on labor market risks, we disregard the old-age component of the SSI program.

[16] In particular, individuals must have income below a "countable income limit", which typically is slightly below the official poverty line (Daly and Burkhauser, 2005). As in the case of Food Stamp eligibility, SSI eligibility also has an asset limit which we disregard (see note 13).

[17] There is a similar argument to be made when replacing low non-health productivity with low job market opportunities (low values of λ).

where qualifying for disability programs is costly, the relationship between economic conditions and program participation is much stronger for permanent than for transitory economic shocks. Further, the opportunity cost of applying is greater if income is higher and those in better health have higher incomes. This decision to apply will depend on assets, age and other characteristics.

Benitez-Silva et al. (2006) characterize in a very compelling way the extent of moral hazard in disability insurance applications. In particular, they show that 40% of recipients do not conform to the criterion of the SSA. This raises the question of whether the "cheaters" are not at all disabled or whether they have only a partial disability. With our characterization of individuals as falling into categories severely restricted ($L = 2$) and at least partially restricted ($L = 0$), we are able to explore this issue.

The criteria quoted above specifies "any substantial gainful activity": this refers to a labour supply issue. However, it does not address the labour demand problem. Of course, if the labour market is competitive this will not be an issue because workers can be paid their marginal product whatever their productivity level. In the presence of imperfections, however, the wage rate associated with a job may be above the disabled individual's marginal productivity. The Americans with Disability Act (1992) tries to address this question but that tackles the issue only for incumbents who become disabled.

2.5 Solution

There is no analytical solution for our model. Instead, the model must be solved numerically, beginning with the terminal condition on assets, and iterating backwards, solving at each age for the value functions conditional on work status. The solution method is discussed in more detail in the Web Appendix. Here we describe the main features of the algorithm used.

We start by constructing the value functions for the individual when employed and when out of work. When employed, the state variables are $\{A_{it}, \varepsilon_{it}, L_{it}\}$, corresponding to current assets, individual productivity and health status. We denote the value function when employed as V^e. When unemployed, there are three alternative discrete states the individual can be: unemployed and not applying for disability (giving a value V^n), unemployed and applying for disability (giving a value V^{App}), and unemployed and already receiving disability insurance (giving a value V^{Succ}).

We describe here the specification of the value function when employed and leave the discussion of the other value functions to the Web Appendix. The value function if working can be written as:

$$V_{it}^{e}\left(A_{it},\varepsilon_{it},L_{it}\right)=$$
$$\max_{c}\left\{\begin{array}{c} U\left(c_{it},P_{it}=1;L_{it}\right)+ \\ \beta\delta E_{t}\left[V_{it+1}^{n}\left(A_{it+1},\varepsilon_{it+1},L_{it+1},DI_{it+1}^{Elig}=1\right)\right] \\ +\beta\left(1-\delta\right)E_{t}\max\left\{\begin{array}{c} V_{it+1}^{n}\left(A_{it+1},\varepsilon_{it+1},L_{it+1},DI_{it+1}^{Elig}=1\right) \\ V_{it+1}^{e}\left(A_{it+1},\varepsilon_{it+1},L_{it+1}\right)\end{array}\right\}\end{array}\right.$$

where DI_{it+1}^{Elig} is an indicator for whether the individual is eligible to apply for DI. Our model has discrete state variables for: Wage productivity, Work limitation status, Participation, Eligibility to apply for DI (if not working), and Length of time on DI (over 1 year or less than 1 year). The only continuous state variable is assets.

Value functions are increasing in assets A_t but they are not necessarily concave, even if we condition on labor market status in t. The non-concavity arises because of changes in labor market status in future periods: the slope of the value function is given by the marginal utility of consumption, but this is not monotonic in the asset stock because consumption can decline as assets increase and expected labor market status in future periods changes. This problem is also discussed in Lentz and Tranaes (2001). By contrast, in Danforth (1979) employment is an absorbing state and so the conditional value function will be concave. Under certainty, the number of kinks in the conditional value function is given by the number of periods of life remaining. If there is enough uncertainty, then changes in work status in the future will be smoothed out leaving the expected value function concave: whether or not an individual will work in $t+1$ at a given A_{it} depends on the realization of shocks in $t+1$. Using uncertainty to avoid non-concavities is analogous to the use of lotteries elsewhere in the literature. In the value functions above, the choice of participation status in $t+1$ is determined by the maximum of the conditional value functions in $t+1$.

2.6 Structural Parameters to Estimate

To summarize, there are four sets of structural parameters that we want to estimate (separately by education). The first set includes parameters characterizing risk: Disability risk (the probability of having a work limitation in t, given past health), the effect of disability on wages (β_1 and β_2), and productivity risk σ_η^2. The second set is labor market frictions: The job destruction rate δ, the arrival rate of job offers when unemployed λ, and the fixed cost of work $F(L)$.The third set of parameters characterize the DI policy parameters: The probability of success in DI application when "young" ($\pi_{L=0}^{Young}, \pi_{L=1}^{Young}, \pi_{L=2}^{Young}$) and when "old" ($\pi_{L=0}^{Old}, \pi_{L=1}^{Old}, \pi_{L=2}^{Old}$), and the probability of reassessment while on DI, P^{Re}. The final set of parameters is preferences: The utility cost of a work limitation θ, the disutility of work η, the coefficient of relative risk aversion γ and the discount rate β. As we will discuss later, some of these parameters will be set to realistic values (taken from the literature) rather than estimated.

3 Identification

Our identification of the unknown parameters proceeds in a number of steps. First, we estimate disability risk directly from transitions between disability states. Second, we estimate the effect of disability on wages using wage data, controlling for selection into work. Third, we estimate productivity risk from unexplained innovations to wages, again controlling for selection into work. Finally, we use indirect inference for the remaining parameters: preferences (the utility cost of disability and the utility cost of participation), labour market frictions, and the parameters that characterize the disability insurance process. To do this, we use a range of auxiliary equations (coefficients from consumption regression, participation over the life-cycle, health status of DI recipients and DI status of individuals of different health).

3.1 Disability Risk

Disability risk is independent of any choices made by individuals in our model, and is also independent of productivity shocks. This means that the disability risk process can be identified structurally without indirect inference. By contrast, the same is not true for the variance of wage shocks which are identified using a selection correction that is based on a reduced form rather than on our structural model. We may include the wage risk parameters in the indirect inference estimation but we do not have to include the disability risk parameters.

3.2 The Wage Process

We modify the wage process (1) to include a measurement error ω_{it}:

$$\ln w_{it} = X_{it}'\alpha + \beta_1 L_{it}^1 + \beta_2 L_{it}^2 + \varepsilon_{it} + \omega_{it} \tag{5}$$

with $\varepsilon_{it} = \varepsilon_{it-1} + \zeta_{it}$ as before. We make the assumption that the two errors ζ_{it} and ω_{it} are independent. Based on evidence from e.g., Bound and Krueger (1995), we assume that the measurement error ω_{it} may be serially correlated (an MA(1) process). Our goal is to identify the variance of the productivity shock σ_η^2 as well as β_1 and β_2. A first complication is selection effects due non-participation. Wages are not observed for non-participants. Moreover, non-participation depends on wages. Finally, non-participation may depend directly on disability shocks as well as the expectation that the individual will apply for DI in the subsequent period (which requires being unemployed in the current period). We observe neither these expectations, nor the decision to apply.

One possible approach is to write a reduced form model of participation:

$$\begin{aligned} P_{it}^* &= X_{it}'\gamma + \delta_1 L_{it}^1 + \delta_2 L_{it}^2 + \theta G_{it} + \vartheta_{it} \qquad (6) \\ &= s_{it} + \vartheta_{it} \end{aligned}$$

where P_{it}^* is the utility from working, and we observe the indicator $P_{it} = \mathbf{1}\{P_{it}^* > 0\}$. Here G_{it} is a vector of exclusion restrictions: They affect the likelihood of observing an individual at work (through an income effect and through affecting the expectation that the individual will apply for DI in the subsequent period), but they do not affect the wage, conditional on X_{it} and L_{it}. We assume that income transfers and an indicator of UI generosity, affecting the opportunity cost of employment, serve as exclusion restrictions. The unobserved "taste for work" ϑ_{it} is correlated with the permanent productivity component ε_{it}. Assume that

$$\begin{pmatrix} \varepsilon_{it} \\ \vartheta_{it} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon\vartheta} \\ & 1 \end{pmatrix} \right)$$

Under these assumptions, the wage for labor market participants is thus:

$$\begin{aligned} E\left(\ln w_{it} | P_{it}^* > 0, X_{it}, L_{it}\right) &= X_{it}'\alpha + \beta_1 L_{it}^1 + \beta_2 L_{it}^2 + E\left(\varepsilon_{it} | P_{it}^* > 0, X_{it}, L_{it}\right) \\ &= X_{it}'\alpha + \beta_1 L_{it}^1 + \beta_2 L_{it}^2 + \sigma_{\varepsilon\vartheta}\lambda\left(s_{it}\right) \end{aligned}$$

assuming no selection on the measurement error. The Mills' ratio term $\lambda(s_{it}) = \frac{\phi(s_{it})}{\Phi(s_{it})}$, where $\phi(.)$ and $\Phi(.)$ denote the p.d.f. and c.d.f. of the standard normal distribution, respectively. Thus, one can estimate

$$\ln w_{it} = X'_{it}\alpha + \beta_1 L^1_{it} + \beta_2 L^2_{it} + \sigma_{\varepsilon\vartheta}\lambda(s_{it}) + v_{it} \tag{7}$$

only on the sample of workers, and with $E(v_{it}|P^*_{it} > 0, X_{it}, L_{it}) = 0$. The resulting estimates of β_1 and β_2 should be interpreted as the estimates of the effect of work limitations on *offered* wages.

3.3 Productivity Risk

To identify the variance of productivity shocks, we define first the "adjusted" error term:

$$g_{it} = \Delta\left(\ln w_{it} - X'_{it}\alpha - \beta L^1_{it} - \beta_2 L^2_{it}\right) \tag{8}$$

From estimation of α, β_1 and β_2 described above we can construct the "adjusted" residuals (8), use them as they were the true "adjusted" error terms (MaCurdy, 1982). Assuming for simplicity of notation that ω_{it} is i.i.d., we can then identify the variance of productivity shocks and the variance of measurement error using the following moment restrictions:

$$\begin{aligned} E(g_{it}|P_{it}=1, P_{it-1}=1) &= \rho_{\zeta\vartheta}\sigma_\eta\lambda(s_{it}) & (9)\\ E\left(g^2_{it}|P_{it}=1, P_{it-1}=1\right) &= \sigma^2_\zeta\left(1-\rho^2_{\zeta\vartheta}s_{it}\lambda(s_{it})\right) + 2\sigma^2_\omega & (10)\\ -E(g_{it}g_{it+1}|P_{it}=1, P_{it-1}=1) &= \sigma^2_\omega & (11)\end{aligned}$$

(see Low, Meghir and Pistaferri, 2006). Here $\rho_{\zeta\vartheta}$ denotes the correlation coefficient between ζ and ϑ (which is not of direct interest). Standard errors are computed with the block bootstrap.

3.4 Preferences and Disability Insurance Parameters

Identification of the remaining structural parameters of interest $(\eta, \theta, \delta, F_{L=0}, F_{L=1}, F_{L=2})$ and the "policy" parameters ($\pi^{Young}_{L=0}, \pi^{Young}_{L=1}, \pi^{Young}_{L=2}, \pi^{Old}_{L=0}, \pi^{Old}_{L=1}, \pi^{Old}_{L=2}$, and P^{Re}) will be achieved by Indirect Inference (see Gourieroux et al, 1993; Smith, 2006). Indirect inference is a simulation-based method that is used when the relevant theoretical moments have no analytical expressions. This is indeed the case for our complex theoretical model. The difference between indirect inference and other methods based on simulations (such as Simulated Method of Moments) is that indirect inference requires only the specification of an approximate model (known as auxiliary model). The auxiliary model is not necessarily the correct data generating process. However, the main idea behind indirect inference is that the parameters of the auxiliary model are related (through a so-called binding function) to the structural parameters of interest. The latter are estimated by minimizing the distance between the parameters of the auxiliary model estimated from the observed data and the parameters of the auxiliary model estimated from the simulated data.

We use the following Indirect Inference auxiliary equations, which overall give us 30 moments: (1) Regression of log consumption on work limitation, disability insurance, participation (and interactions), controlling for a number of other covariates; (2) Participation rates, conditional on disability status and age; (3) Stock of recipients of DI, conditional on disability status and age; and (4) DI status of people of different age and health status.

The Indirect Inference statistical criterion that we use is:

$$\hat{\phi} = \arg\min_{\phi} \left(\widehat{\alpha}^D - S^{-1} \sum_{s=1}^{S} \widehat{\alpha}^S(\phi) \right)' \Omega \left(\widehat{\alpha}^D - S^{-1} \sum_{s=1}^{S} \widehat{\alpha}^S(\phi) \right)$$

where $\widehat{\alpha}^D$ are the moments in the data, $\widehat{\alpha}^S(\phi)$ are the corresponding simulated moments (which are averaged over S simulations) for given parameter values ϕ ($\alpha(\phi)$ is the binding function relating the auxiliary parameters to the structural parameters), and Ω is the weighting matrix. The optimal weighting matrix is the the covariance matrix from the data $\widehat{\Omega}$.[18]

Standard errors of the structural parameters can be computed using the formula provided in Gourieroux et al. (1993), i.e.,

$$var\left(\widehat{\phi}\right) = \left(1 + \frac{1}{S}\right) \left(\frac{\partial \widehat{\alpha}^S(\phi)}{\partial \phi}' \Omega \frac{\partial \widehat{\alpha}^S(\phi)}{\partial \phi} \right)^{-1}$$

If $\dim(\alpha) > \dim(\phi)$, the model generates overidentifying restrictions that can be used to test the model. One can also test for local identification by computing the Jacobian matrix of the binding function and testing whether the matrix has full row rank (see Bond et al., 2007). In what follows we discuss the mapping between structural and auxiliary parameters.

[18] To reduce computational issues, we use $diag\left(\widehat{\Omega}\right)$. We compute standard errors (and the test of overidentifying restrictions) using a formula that adjusts for the use of the non-optimal weighting matrix.

3.4.1 Moments: Consumption Regression

Disability is likely to have two separate effects on consumption: first, disability affects earnings and hence consumption through the budget constraint. The size of this effect will depend on the extent of insurance, both self-insurance and formal insurance mechanisms, such as DI. The extent of insurance from DI obviously depends on being admitted onto the program, but conditional on receiving DI, the extent of insurance is greater for low income individuals because of the progressivity of the system through the AIME and PIA calculation.

The second possible effect of disability on consumption is through non-separabilities in the utility function. For example, if being disabled reduces the marginal utility of consumption (e.g. through loss of appetite) then consumption will fall on disability even if there is full insurance and marginal utility is smoothed over states of disability.

It is important to separate out these two effects. Stephens (2001) calculates the effect of the onset of disability on consumption, but does not distinguish whether the effect is through nonseparability or through the income loss directly.

Our method for separating out these two effects is to use the parameters of the following auxiliary regression:

$$\begin{aligned}\ln c_{it} &= \alpha_0 + \alpha_1 L_{it}^1 + \alpha_2 L_{it}^1 DI_{it} + \alpha_3 L_{it}^2 + \alpha_4 L_{it}^2 DI_{it} + \alpha_5 DI_{it} \\ &\quad + \alpha_6 Y_{it}^P + \alpha_7 t + \alpha_8 t^2 + \alpha_9 A_{it} + \alpha_{10} P_{it} + \upsilon_{it}\end{aligned}$$

The effect of a (severe) disability on consumption for individuals who are not DI insured is given by the parameter α_3. This captures both the income effect and the separability effect. For individuals who are DI insured, the effect of a severe disability on consumption is $(\alpha_3 + \alpha_4)$, and so $(\alpha_3 + \alpha_4)$ captures the preference effect induced by nonseparability.[19] The coefficients α_1 and α_2 correspond to the effects of a moderate disability. We control for permanent income and age because

[19] A heuristic argument for identification is the following. A regression of consumption on work limitation does not identify the non-separability effect because of the presence of busget constraint effects. However, if we could find a group of individuals who are fully insured against disability shocks, then the consumption response to disability can only capture preference effects. Our auxiliary regression is designed to capture this idea through the interaction with the indicator for whether the disabled are insured through the DI program.

we want to compare individuals facing the same level of insurance through the DI system.[20] We control for unearned income to compare individuals with the same potential for self-insurance. The split between α_3 and α_4 is clear when insurance is full. More generally, if insurance is partial, then $(\alpha_3 + \alpha_4)$ captures both the non-separable part and the lack of full insurance for those receiving DI. However, the degree of partial insurance through DI depends on permanent income and age through the AIME formula. Indirect inference exploits this identification intuition without putting a structural interpretation on the values of the α's parameters.

Participation in the labour force can also provide insurance against disability shocks. In addition, participation has a direct effect on the marginal utility of consumption. We use α_{10}, combined with the average participation rates over the life-cycle, to capture this non-separable component and the fixed cost of work.

3.4.2 Moments: Participation over the Life-Cycle

We calculate participation rates by age and by disability status. This is equivalent to run the following auxiliary regression

$$p_{ia}^{L} = \sum_{x=1}^{X} \varphi_x^L \mathbf{1}\left\{age_i \in x, L\right\} + \upsilon_{ia}$$

where p_{ia} is an indicator for whether the person is working, x denote the age bands and there are overall X age bands (we use four 10-year age bands: 23-32, 33-42, etc.). The moments we use as auxiliary parameters are the φ_x^L, that is $X \times L = 12$ auxiliary parameters.

These moments are related to fixed cost of participation with different disabilities, $F(L)$, the utility cost of participation, η, and the labor market frictions. Frictions are identified by average labor market participation and unemployment duration over the life cycle. To see the intuition, consider a world in which there are no food stamps. Because people are born healthy and without assets to finance consumption during unemployment, the decision not to work in the first periods of life is infinitely costly in terms of utility. Hence, if we see people making a transition from employment to unemployment in the first periods of life this must reflect involuntary layoffs, i.e., unemployment rates in the first periods of the life cycle are informative about the job destruction

[20] We construct Y_{it}^P by using the information on individual wages available from entry into the PSID sample until the particular observation at age t.

rate δ (because the reservation asset value is very high at young ages and nobody quits). Finally, the differences in participation across disability status groups is informative about the disability status-specific fixed costs of work (i.e., work is more costly for disabled than for healthy workers), conditioning on wage offers.

3.4.3 Moments: Disability Insurance

There are two ways in which we calculate moments for the stock of DI recipients. First, we consider the composition of DI recipients by health status. This identifies the fraction of recipients who are not truly disabled and helps to pin down the incentive cost. Second, we consider the DI status of individuals within work limitation-types. In particular, we use the fraction of those with a severe limitation who are in receipt of DI. This helps to identify the fraction of the truly disabled who benefit from the insurance and is related to the parameter governing the probability of a successful application. For example, in the data if we observed $x\%$ of the severely disabled receiving DI, then this would suggest x is a lower bound on the probability of acceptance: it would be the fraction of the $L = 2$ on DI if all $L = 2$ individuals applied and no one left the programme. Of course, in practice, the fraction who apply depends on the probability of acceptance and this is why we need to use our model to identify the actual probability of acceptance rather than just taking the observed fractions on DI as the probabilities of acceptance. For both sets of moments, we condition on being younger or older than age 45.

4 Data

We conduct our empirical analysis using longitudinal data from the 1986-1993 Panel Study of Income Dynamics (PSID).[21] The PSID offers repeated, comparable annual data on disability status, disability insurance recipiency, earnings, and food consumption. Its main disadvantage is that the sample of people likely to have access to disability insurance is small and there may be some questions about the variables that define both disability (or work limitation) status and disability insurance status (see below), especially in comparison to the definition of disability of the Social

[21] Due to the retrospective nature of the questions on earnings and consumption, this means our data refer to the 1985-1992 period. We use labor income data before 1985 to construct a measure of permanent income for each individual and each year after 1985.

Security Administration.[22] The PSID sample we use excludes the Latino sub-sample, female heads, and people younger than 23 or older than 62. We also exclude those with missing reports on education, the state of residence, the self-employed, those with less than 3 years of data, and some hourly wage outliers (those with an average hourly wage that is below half the state-level minimum wage and those whose hourly wage declines by more than 75% or grows by more than 400%). Given that the timing of the work limitation question does not coincide with the timing of the DI receipt question (the former refers to the time of the interview, the latter to the previous calendar year), we also lose the first cross-section of data for each individual. The CEX sample we construct to do the imputation of consumption tries to mimic these selections as closely as possible. The CEX only uses families headed by a male, reporting data between 1986 and 1992, with no missing data on the region of residence, aged 23 to 62, not self-employed, reporting data for all interviews (so an annual measure of consumption can be constructed), with complete income response, non-zero consumption or food, and not living in student housing.

Most of the structural analyses of DI errors have used HRS or SIPP data. The HRS has the advantage over the PSID of asking very detailed questions on disability status and DI application, minimizing measurement error and providing a direct (reduced form) way of measuring errors. The SIPP is a much larger data set than the PSID. However, these data sets also have disadvantages. The HRS samples from a population of older workers and retirees (aged above 50). This is an important limitation because the high current levels of DI were associated with sharp increases in the flow-on rates for the under-50s: Figure 1 shows that male workers younger than 40 account for 20 to 25% of new entrants in the Disability Insurance program in recent years, and between 40% and 50% of new entrants are under 50. We use the PSID to understand this behavior because it samples individuals from all ages and follows them across their life-cycle. The SIPP lacks any consumption data. This is problematic because an important element of our model is the state dependence in utility induced by health.

[22] We considered using HRS instead of PSID. The HRS has the advantage over the PSID of asking very detailed questions on disability status and various impairments, minimizing measurement error. However, as said earlier the HRS is a survey of older workers (over 50), so it cannot be used to study behavior at the early stage of the life cycle. Further, DI recipiency status cannot be distinguished from SSI recipiency status. Finally, there is no strict alignment between the timing of the disability questions and the consumption questions. In particular the additional modules on consumption are conducted in 2001 and 2003 but the core questions are asked in 2000 and 2002, and these questions are asked to individuals born before 1947. In future drafts, we will consider complementing our analysis with SIPP data.

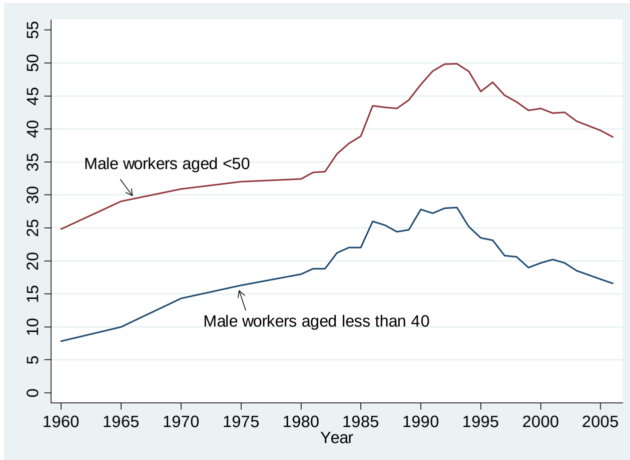

Figure 1: Proportion of new DI awards

Despite its limitation, the PSID matches quite well a number of facts and aggregate statistics. For example, estimates of disability rates in the PSID are similar to those obtained in other, larger data sets (CPS, SIPP, NHIS - and HRS conditioning on age, see Bound and Burkhauser, 1999). Moreover, PSID disability rates by age compare well with aggregate data (see Web Appendix). The match is good also in the time series. In the population, the proportion of people on DI has increased from 2.4% to 4.3% between 1985 and 2005. In the PSID the increase between 1985 and 2005 is from 2.4% to 4.5%.

4.1 Disability Data

We define a discrete indicator of work limitations (L_{it}), based on the following questions: (a) *Do you have any physical or nervous condition that limits the type of work or the amount of work you can do?* To those answering "Yes", the interviewer then asks: (b) *Does this condition keep you from doing some types of work?* The possible answers are: "Yes", "No", or "Can do nothing". Finally, to those who answer "Yes" or "No", the interviewer then asks: *For work you can do, how much does it limit the amount of work you can do?* The possible answers are: "A lot", "Somewhat", "Just a little", or "Not at all".

We distinguish between no work limitations ($L_{it} = 0$), moderate limitations ($L_{it} = 1$) and severe limitations ($L_{it} = 2$). We assume that an individual is affected by moderate work limitations if he answer "Yes" to the first question, "Yes/No" to the second and "Somewhat" to the third (50%), or if he answer "Yes" to the first question, "Yes/No" to the second and "Just a little" to the third (50%). We assume that an individual is affected by severe limitations if he answer "Yes" to the first question and "Can do nothing" to the second question (11%), if he answer "Yes" to the first question, "Yes" to the second and "A lot" to the third (87%), or if he answer "Yes" to the first question, "No" to the second and "A lot" to the third (2%).

The validity of these self-reports is somewhat controversial for two reasons: first, individuals may over-estimate their work limitation in order to justify their disability payments or their non-participation in the labour force. Second, health status may be endogenous, and non-participation in the labour force may affect health (either positively or negatively). Regarding the first criticism, Bound and Burkhauser (1999) survey a number of papers that show that self-reported measures are highly correlated with *clinical* measures of disability (Bound and Burkhauser, 1999). Benitez-Silva et al., (2003) show that self-reports are unbiased predictors of the definition of disability used by the SSA. Burkhauser and Daly (1996) show the validity of the PSID measures using the 1986 health supplement. Finally, Burkhauser et al. (2002) show that the employment trends for working-age men and women found in the CPS and the NHIS based on working limitation definition of disability yield trends in employment rates between 1983 and 1996 that are not significantly different from the employment trends for the broader population of people with an impairment. See however Kreider (1999) and Kreider and Pepper (2007) for evidence based on bound identification that disability is over-reported among the unemployed.

Regarding the second criticism of the endogeneity of health status, Stern (1990) and Bound (1991) both find positive effects of non-participation on health, but the effects are economically small. Further, Smith (2004) finds that income does not affect health once one controls for education.

4.2 Disability Insurance

To identify whether an individual in the PSID is receiving disability insurance, we use a question that asks whether the amount of social security payments received was due to disability.[23] This

[23] The survey first asks the amount of Social Security payments received in year t by the year $t+1$ head. Then, it asks *Was that disability, retirement, survivor's benefits, or what?*. Possible responses are: 1) Disability, 2) Retirement, 3) Survivor's benefits; dependent of deceased recipient, 4) Dependent of disabled recipient, 5) Dependent of retired recipient, 6) Other, 7) Any combination of the codes above.

question is asked from the 1986 wave onwards. Prior to 1986, the question was not targeted to the head of the household, and so we cannot distinguish the recipient of the insurance. Between 1994 and 2003 the questions that allow us to identify who is receiving DI are present in the questionnaire but not in the PSID early release version for those years.[24] Starting with the 2005 wave, the information is released again.

4.3 Consumption Data

One difficulty with the PSID is that the consumption in the data refers only to food. By contrast, in the model, the budget constraint imposes that over the lifetime, all income is spent on (non-durable) consumption. To compare consumption in the model to consumption in the data, we create non-durable consumption in the data by an imputation procedure. We estimate in the CEX the following regression:

$$\ln c_{it} = \sum_{j=0}^{K} \theta_j \left(\ln F_{it}\right)^j + X_{it}'\mu + \varepsilon_{it}$$

We use a third-degree polynomial in $\ln F$ and control for a cubic in age, number of children, family size, dummies for white, education, region, year, a quadratic in log before-tax family income, labor market participation status, an indicator for whether the head is "Ill, disabled, or unable to work", an indicator for whether the head is receiving social security payments (which for workers aged 62 or less should most likely capture DI), and interaction of the disability indicator with log food, log income, a dummy for white, the DI indicator, and a quadratic in age. The R^2 of the regression is 0.79.

We next define in the PSID the imputed value:

$$\widehat{\ln c_{it}} = \sum_{j=0}^{K} \widehat{\theta}_j \left(\ln F_{it}\right)^j + X_{it}'\widehat{\mu}$$

This is the measure of consumption we use in the analysis that follows.

5 Results

Before delving into the details of estimation, it is useful to provide some basic information about the data. Table 1 reports some sample statistics separately for individuals with no limitations ($L = 0$)

[24] This was an oversight. The PSID plans to release the information for 1994-2003.

and for those with moderate ($L = 1$) and severe work limitations ($L = 2$), and by education (using sampling weights throughout). As said earlier, this is a sample of male heads aged 23-62 who are not self-employed (monetary variables are expressed in 1992 dollars). Regardless of education, the disabled are older, less likely to be married or white, with a smaller family, less likely to be working, and more likely to be on DI. Their family income, wages, and food spending are lower, but income from transfers (both private and public transfers) is higher. The high educated have higher participation rates and lower DI recipiency rates.

Table 1: Sample Statistics

	Low Education			*High Education*		
	$L = 0$	$L = 1$	$L = 2$	$L = 0$	$L = 1$	$L = 2$
Age	40.28	44.80	48.81	39.46	42.69	46.07
% Married	0.79	0.84	0.69	0.77	0.72	0.61
% White	0.84	0.90	0.80	0.91	0.92	0.76
Family size	3.01	3.16	2.61	2.92	2.70	2.57
Family income	43,912	39,715	26,416	66,945	51,728	36,098
Income from transfers	1,758	4,667	10,284	1,637	4,700	11,358
% Working now	0.90	0.71	0.15	0.94	0.77	0.44
% Annual wages > 0	0.97	0.81	0.19	0.98	0.89	0.48
Annual hours worked	2,074	1,573	258	2,192	1,820	831
Annual wages	28,709	19,967	2,801	44,979	28,854	13,525
% DI recipient	0.01	0.08	0.52	0.00	0.03	0.31
Food spending	5,352	5,223	4,198	6,232	5,738	5,223
N	9,112	784	635	8,003	415	171

5.1 Disability Risk

Figure 2 plots selected $\Pr(L_{it} = j | L_{it-1} = k)$.[25] Note that for these graphs we relax the age restrictions (23-62) and focus on the entire life-cycle (until age 82). These are transition probabilities that are informative about the "disability risk". For example, $\Pr(L_{it} = 2 | L_{it-1} = 0)$ is the probability that an individual with no work limitations is hit by a shock that places him in the severe work limitations category. Whether this is a persistent or temporary transition can be answered by looking at the value of $\Pr(L_{it} = 2 | L_{it-1} = 2)$.

[25] To obtain these plots, we first construct a variable that equals the mid-point of a 10-age band (23-32, 33-42, etc.). We then regress an indicator for the joint event $\{L_{it} = j, L_{it-1} = k\}$ on a quadratic in the mid-age variable, conditioning on education and the event $\{L_{it-1} = k\}$. The predicted value of this regression is what we plot in the figure.

The top left panel of Figure 2 plots $\Pr(L_{it} = 0|L_{it-1} = 0)$, i.e., the probabilities of staying healthy. This probability declines over the working part of the life cycle from 0.97 to about 0.92 for the high educated and more rapidly, 0.96 to 0.88, for the low educated . The decline is equally absorbed by increasing probabilities of transiting in moderate and severe work limitations. The top right panel plots the latter, $\Pr(L_{it} = 2|L_{it-1} = 0)$. This probability increases over the life cycle, and the increase is faster for the low educated (1% to 7% vs. 1% to 11%). The probability of full recovery following a severe disability declines over the working portion of life-cycle. For the low educated, such probability is consistently below that of the high educated.[26] Finally, the probability of persistent severe work limitations, $\Pr(L_{it} = 2|L_{it-1} = 2)$ increase strongly with age, and more so for those with low education. In sum, the low educated face worse health risk than the high educated group, with higher probabilities of bad shocks occurring, a lower probability of recovering during the work life, and higher persistence of disability shocks. From now on, all the results we report will refer only the subsample of individuals with low education (high school degree or less). For the purposes of this paper, this is the most interesting group to study both because of higher incidence of disability risk and because of high DI participation. We will come back to the separation between the two education groups when we discuss our welfare analysis.

5.2 Wage Process

In Table 2 column (1) we report the results of estimating a simple probit regression for participation. Participation is monotonically decreasing in the degree of work limitations. We report marginal effects. Thus, the interpretation is that among the low educated, the probability of working declines by 0.13 units at the onset of moderate work limitations, and by 0.54 percentage points at the onset of severe work limitations. As for our exclusion restrictions, their signs are correct (higher income from transfers and a more generous welfare system should increase the opportunity cost of work), and the effects are statistically significant.[27] The other effects have signs that are consistent with previous evidence.

[26] The fact that the low educated have higher probability of recovery during retirement may reflect a selection effect.

[27] To obtain a measure of the generosity of the UI program in the state where the worker lives, we rank states according to the maximum weekly UI benefit (which we take from current legislation). Our measure of generosity is the rank variable, which varies over time and across states. Income from transfers is the sum of private and public transfers. We also used a measure that excludes transfers received by the head, and find virtually identical results.

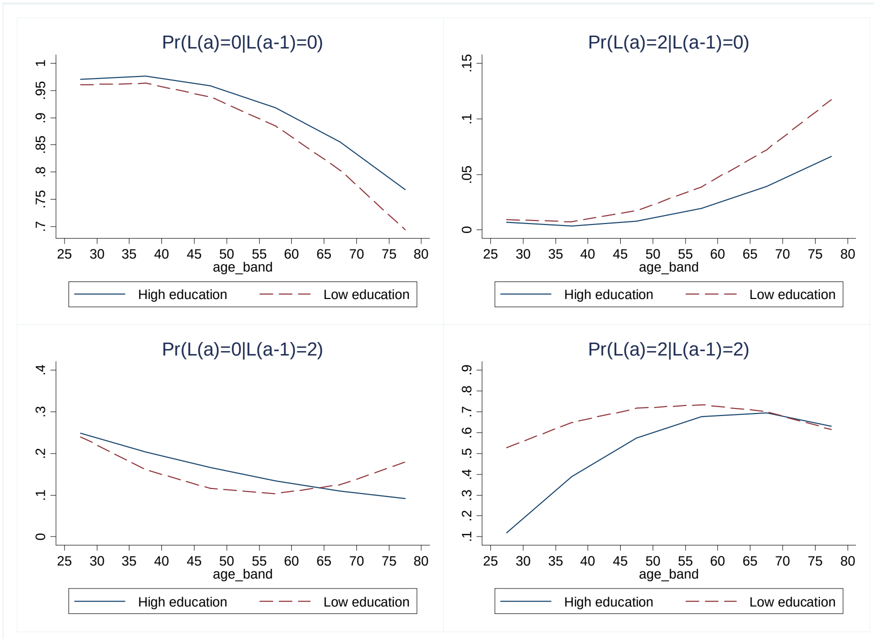

Figure 2: Selected (smoothed) Markov transition probabilities $\Pr\left(L_{ia} = j | L_{ia-1} = k\right)$, by education.

In columns (2) and (3(we report estimates of the log wage process with and without correcting for endogenous selection into work, respectively. This makes a substantial difference. The wage loss associated with the onset of work limitations is higher when selection is being taken into account (+25%). The sign of the Mills' ratio suggests positive selection on unobservables (i.e., people with bad realization of their permanent component quit into unemployment - or are laid-off), and it is statistically significant. The reason for the selection bias is simple. From column (1), an increase in work limitations pushes some people out of work. Those who leave work tend to be those with low unobserved propensity to work (i.e., low π_{it}), which also tend to be individuals with low unobserved permanent income (i.e., low ε_{it}). Hence, if one ignores selection, the wage loss associated with an increase in work limitations appears attenuated by the fact that, among those with work limitations, those who remain at work are higher-than-average permanent income people. Once selection is taken into account, the full loss of disability is revealed.

Table 2: The log wage equation

Variable	Participation equation (1)	Wage w/out selection (2)	Wage with selection (3)
$\{L_{it} = 1\}$	-0.133 (0.015)	-0.196 (0.020)	-0.212 (0.022)
$\{L_{it} = 2\}$	-0.545 (0.026)	-0.323 (0.041)	-0.402 (0.058)
Age	0.007 (0.001)	0.057 (0.004)	0.059 (0.004)
$\frac{Age^2}{100}$	-0.011 (0.002)	-0.058 (0.005)	-0.060 (0.005)
White	0.045 (0.005)	0.254 (0.010)	0.259 (0.010)
Married	0.055 (0.007)	0.149 (0.014)	0.155 (0.014)
Year dummies	Yes	Yes	Yes
UI generosity	-0.0002 (0.0001)	-.-	-.-
$\frac{\text{Income from transfers}}{1000}$	-0.005 (0.0003)	-.-	-.-
Mills ratio	-.-	-.-	0.079 (0.039)
N	10,531	9,542	9,542

5.3 Productivity Risk

We use the residuals of the wage equation to estimate the variance of permanent productivity shocks as well as the variance of transitory shocks (and the MA(1) parameter, which turns out to be statistically insignificant), allowing for endogenous selection into work (expressions (9)-(11)). The results are in Table 3. The numbers are similar to estimates reported elsewhere (see Meghir and Pistaferri, 2005). This suggests that stripping out the variability in wages due to health shocks does not have much impact on the estimates of productivity risk, presumably because disability is a relatively low probability event.

Table 3: The variances of the productivity shocks

Parameter	Estimate
Permanent shock	0.028 (0.009)
Measurement error (Transitory)	0.036 (0.007)
$\rho_{\zeta\vartheta}$	0.468 (0.117)

5.4 Estimates from Indirect Inference

Here we report the estimates we obtain using Indirect Inference. First, we set some parameters to realistic values (see Table 4):

Table 4: Exogenous Parameters

	Value	Frequency
γ	1.5	
R	0.016	Annual
β	0.025	Annual
T^W	200 (40 years)	
T^R	40 (10 years)	
λ	0.73	Quarterly

Ideally, we would identify the value of λ by using durations of unemployment by disability status. However, there are substantial censoring problems, as well as a large amount of noise when we stratify by education and health status, and hence we take the value of λ from Low, Meghir and Pistaferri (2009).

Next, we present results from estimating the auxiliary log consumption equation (using imputed data, as detailed above), see Table 5. Our measure of consumption is per adult equivalent (using the OECD equivalence scale $1 + 0.7\,(A - 1) + 0.5K$, where A is the number of adults and K the number of children in the household). We obtain a good match between data and simulations. The signs and in most cases even the magnitude of the coefficients are similar. These numbers are not intrinsically interesting, however. It is their link with structural parameters that it is more interesting for our purposes.

Table 5: The Log Consumption Equation

Variable	Baseline	Simulations
$\{L_{it} = 1\}$	-0.121 (0.022)	-0.072
$\{L_{it} = 2\}$	-0.184 (0.037)	-0.146
$\{L_{it} = 1\}\,DI$	0.276 (0.105)	0.131
$\{L_{it} = 2\}\,DI$	0.486 (0.094)	0.260
DI	-0.278 (0.083)	-0.008
Employed	0.456 (0.029)	0.337

Controls: Age, Age^2, Unearned income, Permanent income

Table 6 shows participation over the life cycle for people in different work limitation categories. Our simulations match quite well participation of all disability types, but we do not match the full decline in participation with age that is observed in the data.

Table 6: Labor Market Participation by Disability Status

Age band	No limitation		Moderate limitation		Severe limitation	
	Data	Simul.	Data	Simul.	Data	Simul.
23-32	0.98	0.99	0.87	0.96	0.47	0.46
33-42	0.98	0.99	0.88	0.93	0.31	0.38
43-52	0.98	0.97	0.80	0.82	0.21	0.30
53-62	0.88	0.89	0.53	0.64	0.10	0.23

The last piece of evidence comes from matching DI policy moments, i.e., the health status of DI recipient (Panel A) and the DI status of people with different health, separately for younger (age<45) and older workers (age$\geq$ 45) (Panel B) Our model is capable of matching most of the moments with great accuracy. For example, it matches quite closely the proportions of "cheaters" $\mathrm{Fr}(L = 0|DI = 1, t)$ as well as the proportions of workers "insured" by the DI program $\mathrm{Fr}(DI = 1|L = 2, t)$, which are the reduced form equivalents of the incentive/insurance tradeoff.

Table 7: Disability Insurance Moments

Panel A: "Insurance"			Panel B: "Incentives"		
Moment	Data	Simulations	Moment	Data	Simulations
$\mathrm{Fr}(DI = 1\|L = 2, t < 45)$	28.2	27.5	$\mathrm{Fr}(L = 2\|DI = 1, t < 45)$	63.6	65.1
$\mathrm{Fr}(DI = 1\|L = 2, t \geq 45)$	58.5	60.7	$\mathrm{Fr}(L = 2\|DI = 1, t \geq 45)$	73.2	73.5
$\mathrm{Fr}(DI = 1\|L = 1, t < 45)$	5.8	5.7	$\mathrm{Fr}(L = 1\|DI = 1, t < 45)$	22.9	23.0
$\mathrm{Fr}(DI = 1\|L = 1, t \geq 45)$	15.5	14.7	$\mathrm{Fr}(L = 1\|DI = 1, t \geq 45)$	17.0	14.8
$\mathrm{Fr}(DI = 1\|L = 0, t < 45)$	0.23	0.24	$\mathrm{Fr}(L = 0\|DI = 1, t < 45)$	13.6	11.9
$\mathrm{Fr}(DI = 1\|L = 0, t \geq 45)$	1.4	2.2	$\mathrm{Fr}(L = 0\|DI = 1, t \geq 45)$	9.8	11.7

In Table 8 we report the Indirect Inference estimates obtained by minimizing the distance between the moments computed from the data (i.e., those reported in Tables 2, 3, 5, 6, and 7), and the equivalent moments computed from the simulated model. We estimate that a moderate (severe) disability induces about a 4% (8%) loss of utility in terms of consumption. Participation induces a 32% loss.[28] The fixed costs are reported as the fraction of average offered wage income at

[28] An alternative way to estimate the preference parameters η and θ is through a formal Euler equation, using as instruments for the change in disability status and the change in participation past values of the variables. We obtain

age 23. They rise with the degree of disability. We estimate that a job is destroyed on average every 26 quarter. The probability of success of DI application increases with age and disability status. Each DI recipients faces a 5% probability of being re-assessed after the first period on DI. The estimates of the success probabilities by type (age and work limitation status) provide information on the extent of type I and type II errors, which we comment on next.

Table 8: Estimated Parameters

Frictions and Preferences			Disability Insurance Program	
Parameter		Estimate	Parameter	Estimate
θ	Cost of disability	-0.039	$\pi_{L=0}^{Young}$	0.002
η	Cost of part.	-0.32	$\pi_{L=0}^{Old}$	0.009
δ	Job destruction	0.049	$\pi_{L=1}^{Young}$	0.103
$F_{L=0}$	Fixed cost	0.10	$\pi_{L=1}^{Old}$	0.14
$F_{L=1}$	Fixed cost	0.31	$\pi_{L=2}^{Young}$	0.35
$F_{L=2}$	Fixed cost	1.20	$\pi_{L=2}^{Old}$	0.72
			P^{Re}	0.05

Note: Fixed costs are reported as the fraction of average offered wage income at age 23. All parameters significant (using asymptotic standard errors, not correcting for first stage estimates).

6 Implications

Our theoretical framework and structural estimates of the model can be used to study the implications of the existing DI program, as well as to evaluate the welfare effects of modifying the features of the current program.

6.1 Success of the DI Screening Process

One important issue is to evaluate the success rate of the current DI Screening Process. We first look at the Award rate: $\Pr(DI = 1 | DI^{App} = 1)$. We estimate this rate (using our structural model and estimated parameters) to be 0.40. During the period covered by our data (1986-92), there were 3.3 million awards made to 7.8 million applicants, resulting in a 42% success rate.[29] Our estimate

estimates for θ of -0.036 (s.e. 0.060) and for η of -0.597 (s.e. 0.155). The Sargan statistic has a p-value of 66%. The first-stage F-test is 746 for the change in disability and 365 for the change in participation. It is comforting that two different estimation stategies give very similar results for the two parameters of interest (albeit less precise).

[29] See Table 26, Annual Statistical Report on the Social Security Disability Insurance Program, 2000.

contrasts quite well also with the reduced form estimates (0.45) obtained by Bound and Burkhauser (1999) and others using data on DI application and DI receipt from the HRS.

Given that the true disability status of an applicant is private information, SSA evaluators are bound to commit two types of errors: Admitting into the DI program undeserved applicants and rejecting those who are truly disabled. How large are the probabilities associated with these errors? Consider first the extent of false positives (the proportion of healthy individuals who apply receiving DI). From Table 8, these type II errors have probabilities ranging from 0.2% (young non disabled) to 14% (old moderately disabled). Similarly, we can look at the Award Error: $Pr(L = \{0, 1\}|DI = 1, DI^{App} = 1) = 0.10$. In the literature, we have found reduced form estimates that are fairly similar, 0.16 to 0.22 in Benitez-Silva et al. (1999), depending on the statistical assumptions made, and 0.19 in Nagi (1969).

Consider next the probability of false negatives (i.e., the proportion of severely disabled who apply and do not receive DI). From Table 8, we estimate that the type I errors are 65% for the younger and 28% for the older workers. The fraction of those who are rejected who are disabled, the Rejection Error, is given by $Pr(L = 2|DI = 0, DI^{App} = 1) = 0.43$. This is again similar to Benitez-Silva et al. (1999), who report 0.52-0.60, and Nagi (1969), 0.48. These comparisons confirm that our structural model is capable of replicating quite well reduced form estimates obtained using direct information on the application and award process. Our estimated award process is slightly more efficient than previous estimates, but the differences are slight.

Finally, with an estimated reassessment rate of 5%, we predict that an individual on DI is expected to have his disability status reviewed approximately every 20 quarters.[30] To get a gauge of the actual numbers involved, consider that during the fiscal years 1987-1992 (the years covered by our sample) the SSA conducted a total of 1,066,343 Continuing Disability Reviews (CDR). Subtracting from the stock of disabled workers in current payment status the flow of awards for each year, we estimate a probability of re-assessment of 7%, not far from our estimate.

[30] By law, the SSA is expected to perform Continuing Disability Reviews (CDR) every 7 years for individuals with medical improvement not expected, every 3 years for individuals with medical improvement possible, and every 6 to 18 months for individuals with medical improvement expected. In practice, the actual number of CDRs performed is lower.

6.2 Changing Parameters of the DI Process

The most important use of our model is the ability to measure the welfare effects of changing the main parameters of the DI programs. Consider making the program "stricter". In one form or another, this suggestion has been advanced as one possible solution to the "moral hazard" problem. To tackle this issue, one needs to define first a measure of strictness of the program. Suppose that Social Security DI evaluators decide whether to award DI as a function of a signal about the applicant's disability status:

$$S_{it} = \alpha_L^t + \xi_{it}$$

The mean of the signal (α_L^t) varies by age (for simplicity, for two age groups defined by age<45 and age≥45), and by work limitation status L. ξ is a normally distributed error with variance σ_ξ^2. Assume that the Social Security DI evaluators decide to award DI if $S_{it} > \overline{S}$. The parameter can be interpreted as a measure of the strictness of the DI program (ceteris paribus, an increase in $\overline{S}$ reduces the proportion of people admitted into the program). We connect the estimated structural probabilities described above (π_L^t) with the parameters $\overline{S}$, α_L^t, and σ_ξ^2, through the expression:

$$1 - \pi_L^t = \Pr\left(\text{Rejection}|\, t, L, Apply\right) = \Phi\left(\frac{\overline{S} - \alpha_L^t}{\sigma_\xi}\right)$$

where $\Phi\left(.\right)$ is the c.d.f. of the standard normal. Using the 6 probabilities of acceptance (by type and age) from the estimation and using the normalizations $\alpha_{L=2}^{Old} = 1$ (for the group most likely to be admitted into DI) and $\alpha_{L=0}^{Young} = 0$ (for the group least likely to be admitted into DI), we can back up estimates of the threshold $\overline{S}$, α_L^t, and σ_ξ^2 (for $t =$\{"Young", or "Old"\} and $L = \{0, 1, 2\}$). Figure **??** illustrates the extent of some of the errors under the estimated DI program. The area on the left of $\overline{S}$ under the dashed curve (the one labeled $f(S|L = 2, t \geq 45)$) measures the probability of rejecting a deserving DI applicant (type I error, or false negatives). The area on the right of $\overline{S}$ under the dotted curve (labeled $f(S|L = 0, t < 45)$) measures the probability of accepting into the DI program an undeserving DI applicant (type II error). Increasing the strictness of the test (increasing) reduces the probability of false positives (reduces the extent of the incentive problem), but also increases the probability of false negatives (reduces the extent of insurance provided by

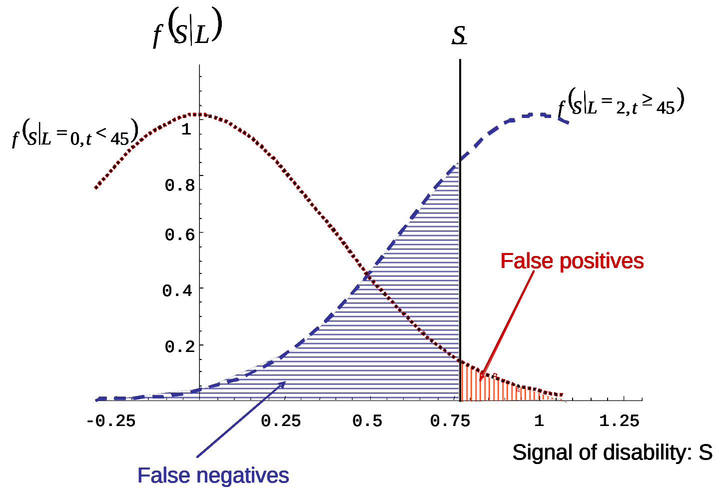

Figure 3: Type I and Type II errors.

the program). Such a policy has both benefits and costs, trading off incentives against insurance, and we use our model to determine which dominates when the strictness of the test changes.[31]

Consider the effect of changes in the acceptance threshold To see whether a change would be welfare improving and if so, in what direction, we consider the following strategy. We hold the government budget constraint constant, which is achieved by adjusting the proportional payroll tax (this is done iteratively because labor supply changes as a consequence). We calculate expected utility and the extent of moral hazard (false applications) for different values of the acceptance threshold $\overline{S}$, as well as other statistics of interest.

Figure 4 reports the results of this experiment. We find that the "optimal" acceptance threshold lies on the left of the estimated (actual) threshold, i.e., decreasing the strictness of the test is welfare-improving (note that, absent a theoretical framework, nothing could be said about whether the optimal threshold is higher or lower than the estimated one). In the optimal scenario (obtained by maximization of expected utility *behind the veil of ignorance*, i.e., before individuals discover their types etc.), the acceptance threshold is about 30% higher than the estimated one. Hence, we find that it is welfare enhancing to make the medical test less strict to reduce false negatives (and the rejection error), despite the worsening in the degree of moral hazard. Part of the explanation

[31] One alternative is to invest in technologies that increase the degree of information about individuals' true disability status (i.e., reducing σ_ξ^2). Unfortunately, this policy is typically very expensive or unfeasible.

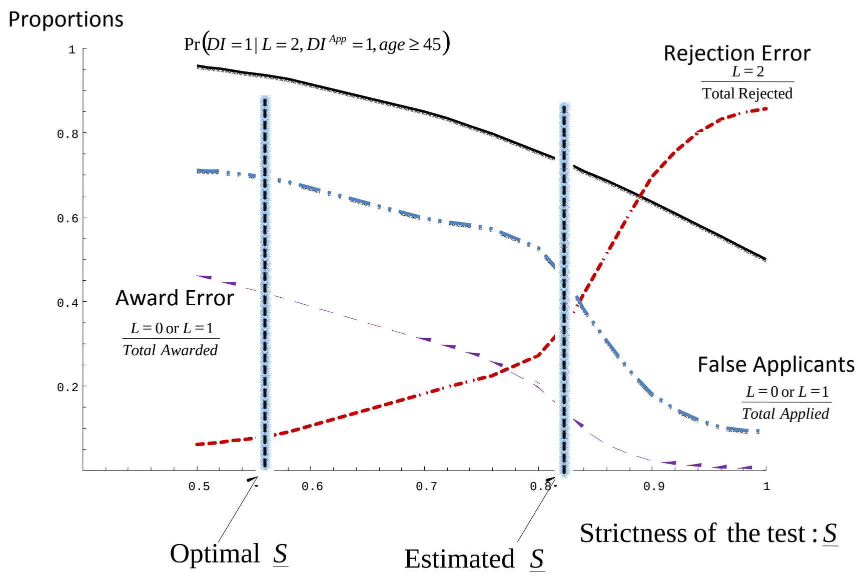

Figure 4: Optimal strictness.

is that type I error is particularly high for the young, who have very few assets to substitute social insurance with self-insurance (both because they have had only a few years to accumulate any wealth and because precautionary saving against a low probability event is very inefficient).[32]

We also consider changing other parameters of the DI program: (1) Reducing the generosity of payments; (2) Increasing the reassessment rate; and (3) Reducing food stamps benefits [TBW]

7 Conclusions

[TBW].

[32]This result is obtained balancing the government budget within the low education group. If we had a single government budget constraint, with the high educated subsidizing the higher DI recipiency rates of the low educated, our result would be most likely similar (and if anything stronger, because in the scenario we have considered there is some resistance to expand insurance given that taxes must rise to finance the expansion of the DI program - there would be less resistance if the higher taxes were paid by the high educated).

References

[1] Acemoglu, D. and J. D. Angrist (2001), "Consequences of Employment Protection? The Case of the Americans with Disabilities Act", The Journal of Political Economy, Vol. 109, No. 5, pp. 915-957

[2] Adda, J., Banks, J. and H-M von Gaudecker (2007) "The impact of income shocks on health: evidence from cohort data" Institute for Fiscal Studies Working Paper 07/05

[3] Autor, David H. and Mark G. Duggan (2003), "The Rise in the Disability Rolls and the Decline in Unemployment", Quarterly Journal of Economics 118: 157-205.

[4] Autor, David H. and Mark G. Duggan (2006), "The Growth in the Social Security Disability Rolls: A Fiscal Crisis Unfolding", Journal of Economic Perspectives, 20(3), Summer: 71 – 96.

[5] Benitez-Silva, Hugo, Moshe Buchinsky, Hiu Man Chan, Sofia Cheidvasser, and John Rust (2004), "How Large Is the Bias in Self-Reported Disability?", Journal of Applied Econometrics, Vol. 19 (6), 649-670.

[6] Benitez-Silva, Hugo, Moshe Buchinsky, Hiu Man Chan, Sofia Cheidvasser, and John Rust (1999), "An Empirical Analysis of the Social Security Disability Application, Appeal and Award Process", Labour Economics 6 147-178.

[7] Benitez-Silva, Hugo, Moshe Buchinsky, and John Rust (2006), "How Large are the Classification Errors in the Social Security Disability Award Process?", NBER Working Paper 10219.

[8] Black, D., K. Daniel and S. Sanders (2002), "The Impact of Economic Conditions on Participation in Disability Programs: Evidence from the Coal Boom and Bust", American Economic Review 92(1), 27-50.

[9] Bound, John (1989), "The Health and Earnings of Rejected Disability Insurance Applicants", American Economic Review 79: 482 – 503.

[10] Bound, John (1991), "Self-Reported Versus Objective Measures of Health In Retirement Models", Journal of Human Resources 26: 106-38.

[11] Bound, John and Richard V. Burkhauser. “Economic Analysis of Transfer Programs Targeted on People with Disabilities” (1999), in Orley C. Ashenfelter and David Card (eds.), Handbook of Labor Economics. Volume 3C. Amsterdam: Elsevier Science, pp. 3417-3528.

[12] Bound, J., Cullen, J. B., Nichols, A. and L. Schmidt (2004) “The welfare implications of increasing disability insurance benefit generosity” Journal of Public Economics 88:2487-2514

[13] Burkhauser, Richard V. and Mary C. Daly. (1996) "Employment and Economic Well-Being Following the Onset of a Disability," In Disability, Work, and Cash Benefits, edited by Jerry Mashaw, Virginia Reno, Richard Burkhauser, and Monroe Berkowitz, Upjohn Institute for Employment Research, Kalamazoo, MI.

[14] Chen, S. and W. van der Klaauw (2005) “The work disincentive effects of the disability insurance program in the 1990s”

[15] DeLeire, Thomas (2000), “The Wage and Employment Effects of the Americans with Disabilities Act”, Journal of Human Resources 35(4):693-715.

[16] Diamond, Peter and Eytan Sheshinski (1995), "Economic aspects of optimal disability benefits", Journal of Public Econornics 57 (1): 1-23.

[17] Finkelstein, A., E. Luttmer and M. Notowidigdo (2008), "What Good Is Wealth Without Health? The Effect of Health on the Marginal Utility of Consumption", June 2008, NBER Working Paper 14089.

[18] Gallipoli,G. and L. Turner (2009) “Household responses to individual shocks: disability, labour supply and marriage” University of British Columbia, mimeo

[19] Gastwirth, Joseph L. (1972), "On the decline of male labor force participation", Monthly Labor Review 95 (10): 44˜46.

[20] Golosov, Mikhail and Aleh Tsyvinski (2006), "Designing Optimal Disability Insurance: A Case for Asset Testing", Journal of Political Economy 114, 257-279.

[21] Gourieroux, Christian, Alain Monfort, and Eric Renault (1993), “Indirect Inference”, Journal of Applied Econometrics, Vol. 8, Supplement: Special Issue on Econometric Inference Using Simulation Techniques, pp. S85-S118.

[22] Gruber, Jonathan (2000), “Disability Insurance Benefits and Labor Supply”, Journal of Political Economy 108: 1162 – 1183.

[23] Heckman, James J. (1979), “Sample Selection Bias as a Specification Error”, Econometrica 47(1): 153-161.

[24] Hoynes, Hilary Williamson and Robert Moffitt (1997), "Tax rates and work incentives in the Social Security Disability Insurance program: current law and alternative reforms", Working paper no. 6058 (NBER, Cambridge, MA).

[25] Hubbard, Ronal G., Jonathan Skinner and Steven P. Zeldes (1995) “Precautionary Saving and Social Insurance”, Journal of Political Economy 103: 360-399.

[26] Kreider, Brent (1999), “Latent Work Disability and Reporting Bias,” Journal of Human Resources, 734-769.

[27] Kreider, Brent and John Pepper (2007), “Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors,” Journal of the American Statistical Association, June 2007, 432-41.

[28] Low, Hamish, Costas Meghir and Luigi Pistaferri (2006), “Wage risk and employment risk over the life cycle”, American Economic Review (forthcoming).

[29] Meyer, Bruce D. and Wallace K.C. Mok (2007) “Disability, earnings, income and consumption” University of Chicago, mimeo

[30] Nagi, S. Z. (1969), Disability and Rehabilitation. Columbus, OH: Ohio State University Press.

[31] Parsons, Donald O. (1980), “The Decline in Male Labor Force Participation”, The Journal of Political Economy 88, No. 1: 117-134

[32] Rust, John, Moshe Buchinsky, and Hugo Benitez-Silva (2002), "Dynamic Models of Retirement and Disability", Working Paper.

[33] Smith, J. (2004) "Unravelling the SES health connection", Institute for Fiscal Studies Working Paper 04/02.

[34] Smith, Anthony A. Jr. (2006), "Indirect Inference", forthcoming in The New Palgrave Dictionary of Economics, 2nd Edition.

[35] Stephens, Mel (2001), " The Long-Run Consumption Effects of Earnings Shocks", The Review of Economics and Statistics, vol.83, n.1, p.28-36.

[36] Stern, Steven (1989), "Measuring the Effect of Disability on Labor Force Participation." Journal of Human Resources 24 (3, Summer), pp. 361-95.

[37] Waidman, Timothy, John Bound and Austin Nichols (2003), "Disability Benefits as Social Insurance: Tradeoffs Between Screening Stringency and Benefit Generosity in Optimal Program Design," Working Papers wp042, University of Michigan, Michigan Retirement Research Center.