Disability Risk and the Value of Disability Insurance*

Hamish Low
University of Cambridge

Luigi Pistaferri
Stanford University

January 25, 2011

Abstract

We provide a lifecycle framework for comparing the insurance value of disability benefits and the incentive cost. We estimate the risks that individuals face and the parameters governing the disability insurance program using longitudinal US data on consumption, health, disability insurance, and wages. We characterize the economic effects of disability insurance and study how policy reforms impact behavior and household welfare. Because of high rejections rates of disabled applicants, welfare increases as the program becomes less strict, despite the worsening of incentives. False applications decline when reducing generosity and increasing reassessments; these policies increase welfare, despite the decline in insurance.

JEL Codes: D91, H53, H55, J26

Keywords: disability, social security, savings behavior, wage risk

*Low thanks funding from the ESRC as a Research Fellow, grant number RES-063-27-0211. 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 Tom Crossley, Pramila Krishnan, Francesco Lippi, Costas Meghir and Aleh Tsyvinski for detailed comments, and to Katja Kaufmann and Itay Saporta for research assistance. All errors are our own.

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 work limiting 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] and 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 Duggan and Imberman, 2009). The key questions in thinking about the size and growth of the program are whether program claimants are genuinely unable to work, whether those in need are receiving insurance, and how valuable is the insurance provided.

Our paper makes three contributions to the literature on DI. First, we propose a theoretical life cycle framework that allows us to study the effect of disability risk on decisions about labor supply, savings and applying for DI. 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 or work limitation shock which reduces the ability to work. The distinction between the two types of shock to wages is crucial for understanding the incentive problem with the DI program: 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. Since disability status is only imperfectly observable, either type may apply for DI benefits.

Second, we estimate the parameters of this model using microeconomic data. We use PSID data on wages, indicators of disability and its severity, receipt of DI, consumption and employment status to help identify the wage and health risks that individuals face, their preferences and the parameters of the DI program. Almost half of the inflow onto DI comes from those aged under 50, and the use of the PSID enables us to study behavior across the whole life-cycle. In contrast, studies that use the HRS are restricted to those aged over 50. Our estimates show that there are substantial false rejections 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 analyse the impact on welfare and behavior of varying the key policy parameters. We focus on addressing how well insured are individuals against disability risk, how responsive are the number of

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

false applications, labour supply and savings to changes in the details of the DI program, and what are the implications for welfare.[2] The ability to evaluate these questions in a coherent, unified framework is one of the main benefits of the paper. We conduct counterfactual experiments by changing: (a) the generosity of disability payments, (b) the stringency of the screening process, (c) the re-assessment rate, (d) the generosity of alternative social insurance programs. One striking finding of our paper is that the high fraction of false rejections (Type I errors) associated with the screening process of the disability insurance program leads to welfare increasing when the program becomes less strict, despite the increase in false applications. This is because coverage among those most in need goes up. On the other hand, welfare is higher if the generosity of DI is cut and if reassessment is more frequent: both of these reforms have a large impact reducing the number of applications from those with only a moderate disability at little cost in terms of reduced coverage for those in need. It is this distinction between moderate and severe disability which turns out to be particularly important and which is novel in our set-up: we show the extent to which the moderately disabled are more responsive to incentives than the severely disabled, and it is this difference which underlies our policy conclusions.

The issues raised in this paper relate to the literatures on the incentives to make a false application for DI and the disincentive effects of DI on labor supply. It also relates to a smaller literature on the costs of disability shocks. Since disability status is private information, DI evaluators make two types of errors: awarding benefits to undeserving applicants, and denying them to truly disabled individuals. The only direct attempt to measure such errors is Nagi (1969), who uses a sample of 2,454 initial disability determinations. These individuals were examined by an independent medical and social team. Nagi (1969) concluded that about 19% of those initially awarded benefits were undeserving, and 48% of those denied were truly disabled. More recently, Benitez-Silva et al. (2006a) using HRS self-reported disability data on the over 50s, conclude that over 40% of recipients of DI are not truly work limited and this adds to the picture of an inefficient insurance program

The incentive to make a false application rather than to work has been addressed by asking how many DI recipients would be in the labor force in the absence of the program. Identifying an appropriate control group has been controversial (see Parsons, 1980; Bound, 1989). Bound (1989) uses DI applicants who were rejected as his control group. He finds that only 1/3 to 1/2 of rejected applicants are working, and this is taken as an upper

[2] To show the effect on household welfare, we calculate the consumption equivalent to keep expected utility at the start of life constant as policy changes.

bound of how many DI beneficiaries would be working in the absence of the program These estimates have recently been confirmed by Chen and van der Klaauw (2008). Relatedly, Kreider (1999) finds that although DI has important disincentive effects on labour supply, the effects due to changes in DI generosity are not large enough to explain fully the fall in labour force participation. The underlying assumption of these papers is that those who would be working in the absence of DI are false applicants. To tackle false applications directly, Golosov and Tsyvinski (2006) propose introducing an asset test for recipients of DI because those who anticipate making a false claim will accumulate assets to help smooth consumption.

Evidence on the effectiveness of incentives to move the healthy off DI is weak: Hoynes and Moffitt (1997) conclude via simulations that some of the reforms aimed at allowing DI beneficiaries to keep more of their earnings on returning to work are unlikely to be successful and may, if anything, increase the number of people applying for DI. In a similar vein, Acemoglu and Angrist (2001) and DeLeire (2000) examine the effect of the Americans 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. Benitez-Silva et al. (2006b) evaluate the effectiveness of a "$1 for $2 offset" policy, which consists of reducing DI benefits by $1 for every $2 of earnings above a certain level. They find that the policy encourages work by DI beneficiaries, but also encourages entry into the program by individuals attracted by the greater generosity who would not have applied otherwise.

There have been some recent papers identifying the extent of health risk which underlie the need for a DI program. DeNardi et al. (2010) 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 et al. (2007) estimate the effect of income shocks on health and find only small effects. Meyer and Mok (2007) and Stephens (2001) have estimated in a reduced form way the effect of disability shocks on household consumption. Gallipoli and Turner (2009) explore in a structural model the effect of disability shocks on consumption and labor supply.

More generally, however, the broader issue of the value of DI and the effects of DI reform requires combining estimates of the risk associated with health shocks in a framework that allows the evaluation of the insurance and incentives provided by DI. Previous work by Bound et al. (2004, 2010), 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.[3] Our work builds on these papers but extends them by modelling explicitly the joint decision over whether to apply for DI and whether to work at different ages across the life-cycle. Further, we have an explicit measure of disability risk which allows for moderate as well as severe disability, rather than just a disabled/not-disabled split. Finally, we allow for a more flexible specification of the wage process and of preferences, and the addition of labor market frictions and interactions with other social insurance programs. None of these elements are purely cosmetic: we believe this is a necessary set-up to provide enough realism to capture the trade-offs inherent in the DI system. For example, negative productivity shocks unrelated to health (such as shocks to skill prices), as well as a possible lack of employment opportunities, are at the root of the incentive problem - both reduce the opportunity cost of applying for DI independent of health status - and so we need such shocks alongside the risk of work-limiting disabilities if we want to explain the decision to apply for DI when not disabled. As another example, 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, and on what alternative mechanisms of insurance exist.

The rest of the paper is structured 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 summarizes the data used in the estimation of the model, focusing on the data on disability status and on consumption. Section 4 discusses the identification strategy, presents the estimates of the structural parameters and discusses the efficiency of the existing DI system. Section 5 carries out counter-factual policy experiments, reporting the effects on behavior and average household welfare of potential reforms of DI. Section 6 concludes.

2 Life-Cycle Model

2.1 Individual Problem

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

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

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

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 indicator, c_t consumption, and L_t a discrete work limitation (disability) status indicator $\{0,1,2\}$, corresponding to no limitation, a moderate limitation and a severe limitation, respectively. Work limitation status is often characterised by a $\{0,1\}$ indicator (as in Benitez-Silva et al., 2006a). We use a three state indicator to investigate the importance of distinguishing between moderate and severe work limitations. 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

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

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,[4] B unemployment benefits, W the monetary value of the means tested welfare payment, DI the amount of disability insurance payments obtained, SSI the amount of Supplemental Security Income (SSI) benefits, and E^{DI}, E^{UI} and E^W are recipiency $\{0,1\}$ indicators for disability insurance, the means-tested welfare program and unemployment insurance, respectively.[5] Unemployment insurance is paid only on job destruction and only for one quarter; the means-tested welfare program is an anti-poverty program providing a floor to income, similar to food stamps. Both programs are described fully in the online appendix.

The worker's problem is to decide whether to work or not. When unemployed, the decision is whether to accept a job that may have been offered or wait longer. The unemployed person will also have the option to apply for disability insurance (if eligible). Whether em-

[4] That disabled individuals face direct costs of work is recognized explicitly by the Social Security Administration (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).

[5] We do not have an SSI recipiency indicator because that is a combination of receiving DI and being eligible for means-tested transfers.

ployed or not, the individual has to decide how much to save and consume. Accumulated savings can be used to finance consumption at any time, particularly during spells out of work and retirement.

We assume that individuals are unable to borrow: $A_{it} \geq 0$. In practice, this constraint has bite because it precludes borrowing against social insurance and means-tested programs. At retirement, people collect social security benefits which are paid according to a formula similar to the one we observe in reality, and is the same as the one used for DI benefits (see below). Social security benefits, along with assets that people have voluntarily accumulated over their working years, are used to finance consumption during retirement. The structure of the individual's problem is similar to life-cycle models of savings and labour supply, such as Low et al. (2010). The innovations in our set-up are to consider the risk that arises from disability shocks that cause a work limitation, the explicit modelling of disability insurance, and the interaction of disability insurance with other social insurance programs.

There are important differences by skill level both in terms of probability of disability shocks and disability insurance recipiency rates. In particular, if we proxy skill level by 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 after that age the difference increases dramatically. 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 (see the online appendix for more details). To account for these differences, in what follows we assume that all the parameters of the model are education-specific, and much of our focus will be on the low educated. To simplify notation, we omit subscripts defining the skill group of interest.

There are three key elements of the problem: (a) preferences, (b) wages, and (c) social insurance. We discuss these in turn.

2.2 Preferences

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 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] [7]

To interpret θ, consider the full insurance case in which individuals attempt to keep marginal utility constant across states. $\theta < 0$ implies that individuals who are fully insured want more expenditure allocated to the "disability" state, for example because they have larger spending needs when disabled (alternative transportation services, domestic services, etc.). In general, whether consumption is larger or smaller when disabled and fully insured is an empirical question.

2.3 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 work limitation status, as well as the contribution of observable characteristics X_{it}:

$$\ln w_{it} = X_{it}'\alpha + \sum_{j=1}^{2} \varphi_j L_{it}^j + \varepsilon_{it} \tag{1}$$

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

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

The work limitation status of an individual, L_{it}, evolves according to a three state first-order Markov process. 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] We interpret ε_{it} as a measure of individual unobserved productivity that is independent of health shocks - examples would include shocks to the value and price of individual skills.

[6] Lillard and Weiss (1997) also find evidence for $\theta < 0$ using HRS savings and health status data. See Finkelstein et al. (2008) for a recent attempt to measure the effect of health status on the marginal utility of consumption using measures of subjective well-being as a proxy for utility.

[7] In addition to the non-separable effect of disability, there may be an additive utility loss associated with disability. Since disability is not a choice, we cannot identify this additive term. Further, such an additive utility loss would be uninsurable because only consumption can be substituted across states.

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

Equation (1) determines the evolution of individual productivity. Productivity determines the offered wage when individuals receive a job offer. The choice about whether or not to accept an offered wage depends in part 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 labor 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.

2.4 Social Insurance

2.4.1 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 persistent 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 award of disability insurance depends on the following conditions: (1) An individual has to have filed an application; (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 which has lasted, or can be expected to last, for a continuous period of*

[9] Besides SSDI, about 25% of workers in the private sector are also covered by employer-sponsored long-term disability insurance plans. Individuals may also qualify for workers' compensation if the disability is job-related.

at least 12 months."[10]

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, \$1000 a month for non-blind individuals as of 2010), the SSA determine whether the individual has a medical disability that is severe and persistent (per the definition above).[11] If such disability is a listed impairment, the individual is awarded benefits without further review.[12] If the applicant's disability does not match 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 opportunity criterion. Two individuals with identical work limitation disabilities 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 screening process. First, we require that the individuals make the choice to apply for benefits. Second, individuals have to have been at work for at least the 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

[10] Despite this formal criterion changing very little, there have been large fluctuations over time in the award rates: for example, award rates fell from 48.8% to 33.3% between 1975 and 1980, but then rose again quickly in 1984, when eligibility criteria were liberalized, and an applicant's own physician reports were used to determine eligibility. 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.

[11] 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 (1990) tries to address this question but that tackles the issue only for incumbents who become disabled.

[12] 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).

status, age, and education:

$$\Pr\left(DI_{it}=1|\,DI_{it}^{App}=1,L_{it},t\right)=\begin{cases}\pi_L^{Young} & \text{if } t<45\\ \pi_L^{Old} & \text{if } 45\le t\le 62\end{cases} \tag{2}$$

In our model, the expected persistence of the work limitation is captured by the Markov process assumption and is age dependent. Since the SSDI program imposes a severity and persistence requirement on the work limitation, this age dependence of the Markov process is the reason why we make the probability of a successful application for DI dependent on age.[13] Eligibility does not depend on whether an individual quits or the job is destroyed.

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 (based on (2)). We hence depart from the standard assumption made in the literature that DI is an absorbing state because we want to be able to evaluate policies that create incentives for DI beneficiaries to leave the program. 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, where again the probability of success is determined by (2).

SSDI benefits are calculated in essentially the same fashion as Social Security retirement benefits. 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). Benefits are independent of the extent of the work limitation, but are progressive.[14]

We set the value of the benefits according to the actual schedule in the US program. The value of disability insurance is given by

$$D_{it}=\begin{cases}0.9\times\overline{w}_i & \text{if } \overline{w}_i\le a_1\\ 0.9\times a_1+0.32\times(\overline{w}_i-a_1) & \text{if } a_1<\overline{w}_i\le a_2\\ 0.9\times a_1+0.32\times(a_2-a_1)+0.15\times(\overline{w}_i-a_2) & \text{if } a_2<\overline{w}_i\le a_3\\ 0.9\times a_1+0.32\times(a_2-a_1)+0.15\,(a_3-a_2) & \text{if } \overline{w}_i>a_3\end{cases} \tag{3}$$

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

[14] Caps on the amount that individuals pay into the DI system as well as the nature of the formula determining benefits (see equation 3) make the system progressive. Because of the progressivity of the benefits and because individuals receiving SSDI also receive Medicare benefits after two years, the replacement rates are substantially higher for workers with low earnings and those without employer-provided health insurance.

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.[15] We assume $\overline{w}_i$ can be approximated by the value of the permanent wage.

To understand our characterization of the trade-off between genuine applicants and non-genuine applicants, 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 exceeds a certain stringency threshold, $S_{it} > \overline{S}$. Some individuals whose actual disability is less severe than $\overline{S}$ may nonetheless wish to apply for DI if their productivity is sufficiently low because the government only observes S_{it}, a noisy measure of the true disability status. In contrast, 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. Given the opportunity cost of applying for DI, these considerations suggest that applicants will be predominantly low productivity individuals or those with severe work limitations (see Black et al., 2004, for a related discussion).

Benitez-Silva et al. (2006a) characterize in a compelling way the extent of false claimants 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 ($L = 1$), we are able to explore this issue.

2.4.2 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 income. The definition of disability in the SSI program is identical to the one for the DI program The definition of low income is similar to the one used for the Food Stamps program[16] We assume that SSI generosity is identical to the means-tested program

[15] 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.

[16] In particular, individuals must have income below a "countable income limit", which typically is slightly below the official poverty line (Daly and Burkhauser, 2003). As in the case of Food Stamp eligibility, SSI eligibility also has an asset limit which we disregard.

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 detail in the online appendix, which also provides the code to solve and simulate the model.

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 (φ_1 and φ_2 in equation (1)), and productivity risk σ_ζ^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 of a 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 Data

We conduct our empirical analysis using longitudinal data from the 1986-1993 Panel Study of Income Dynamics (PSID).[17] 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 the disability (or work limitation) status of an individual, especially in comparison to the definition of disability of the Social Security Administration. Nevertheless, the PSID matches quite well a number of facts

[17] 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. We do not use data before 1985 because major reforms in the DI screening process were implemented in 1984 (see Duggan and Imberman, 2009). We are unable to use more recent data because between 1993 and 2005 we do not have details on which household member receives DI, although such degree of detail may be available in future releases of the data set.

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 insurance rates by age compare well with aggregate data (see the online appendix), and 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%.

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%).[18] 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.

Disability Data We define a discrete indicator of work limitation status (L_{it}), based on the following questions: (1) *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: (2) *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: (3) *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 use answers to these questions to distinguish between having no work limitation ($L_{it} = 0$), a moderate limitation ($L_{it} = 1$) and a severe limitation ($L_{it} = 2$). We assume that those without a work limitation either answer "No" to the first question or "Not at all" to the third question. Of those that answer "Yes" to the first question, we classify as severely limited those who answer question 2 that they "can do nothing" and those that answer question 3 that they are limited "a lot". The rest have a moderate limitation: their answer to question 3 is that they are limited either "somewhat" or "just a little". This distinction between severe and moderate disability enables us to target our measure of work limitation more closely to that intended by the SSA. In particular, we interpret the SSA criterion as

[18] The hourly wage is defined as annual earnings/annual hours.

intending DI for the severely work limited rather than the moderately work limited.

The validity of these self-reports is somewhat controversial for three reasons: first, subjective reports may be poorly correlated with objective measures of disability; second, individuals may over-estimate their work limitation in order to justify their disability payments or their non-participation in the labour force; third, 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. We provide additional evidence in support of our self-reported measure of work limitation in Table 1 in the online appendix. As in Burkhauser and Daly (1996), we use the 1986 PSID health supplement to show how objective measures of limitation vary with self-reported status, but we distinguish between reports of severe and moderate limitations. The table shows that our disability indicators are significantly correlated with a wide variety of objective measures of impairments, including Activities of Daily Laiving (ADL) variables, hospital stays, extreme BMI values, and mortality.

As for the second criticism, Benitez-Silva et al. (2004) show that self-reports are unbiased predictors of the definition of disability used by the SSA. In other words, there is no evidence that, for the sample of DI applicants, average disability rates as measured from the self-reports are significantly higher than disability rates as measured from the SSA final decision rules. Burkhauser and Daly (1996) show that the employment trends for working-age men and women found in the CPS and the NHIS based on a work limitation definition of disability yields 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. However Kreider (1999) provides evidence based on bound identification that disability is over-reported among the unemployed.

Regarding the final 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.

Disability Insurance To identify whether an individual in the PSID is receiving DI, we use a question that asks whether the amount of social security payments received was due to disability.[19]

Consumption Data Consumption in the PSID 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 obtain non-durable consumption in the data with an imputation procedure that uses a regression for nondurable consumption estimated with Consumer Expenditure Survey (CEX) data. The imputation procedure is described in detail in the online appendix.

Sample Statistics Details of the sample used are discussed in Table 2 of the online appendix. 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.

Most of the structural analyses of DI errors have used HRS or SIPP data. Benitez-Silva et al. (2004) use the HRS, which 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. However, there are two important limitations of the HRS: first, the HRS samples only from a population of older workers and retirees (aged above 50). This matters because the high current levels of DI were associated with sharp increases in the inflow rates for the under-50s: 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 (as detailed in the online appendix). The second limitation is that the HRS asks questions about application to DI only to those individuals who have reported having a work limitation at some stage in their life. We use the PSID because it samples individuals from all ages and follows them across their life-cycle and because it asks all individuals specifically about the receipt of disability insurance (rather than social security more generally). The SIPP has the advantage over the PSID of being a much larger data set, but it lacks any consumption data. This is problematic

[19] 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.

because an important element of our model is the state dependence in utility induced by health.

4 Identification and Results

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, labour market frictions, and the parameters that characterize the disability insurance process. To do this, we use a range of auxiliary equations: coefficients from a consumption regression, participation over the life-cycle, health status of DI recipients and the DI status of individuals of different health.

4.1 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: because wages are observed only for workers, wage shocks are identified using a selection correction.

Figure 1 plots selected estimates of the transition probabilities $\Pr(L_{it} = j | L_{it-1} = k)$.[20] These estimates are informative about "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 $\Pr(L_{it} = 2 | L_{it-1} = 2)$.

The top left panel of Figure 1 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

[20] 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.

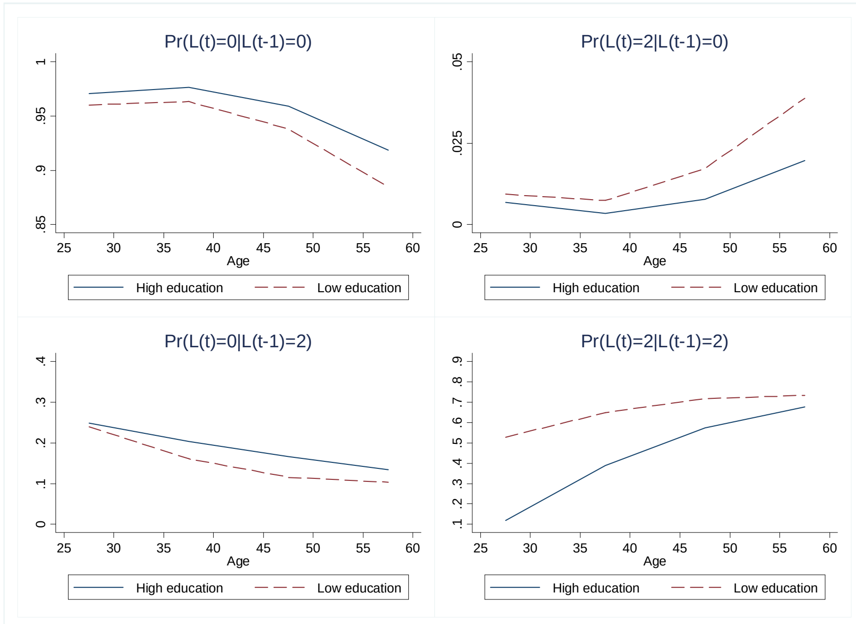

Figure 1: Selected (smoothed) Markov transition probabilities $\Pr(L_{it} = j | L_{it-1} = k)$, by education.

limitations. The top right panel plots the latter, $\Pr(L_{it} = 2 | L_{it-1} = 0)$. This probability increases over the working life, and the increase is faster for the low educated (rising from 1% to 4% vs. 1% to 2%). The probability of full recovery following a severe disability (shown in the bottom left panel) declines over the life-cycle. For the low educated, such probability is consistently below that of the high educated. Finally, the probability of persistent severe work limitations, $\Pr(L_{it} = 2 | L_{it-1} = 2)$ (bottom right panel) increases 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 and a lower probability of recovery.

These differences across education, alongside the much greater prevalence of DI among the low educated, are the reasons why we focus our remaining analysis on the subsample of individuals with low education.

4.2 The Wage Process

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

$$\ln w_{it} = X_{it}'\alpha + \sum_{j=1}^{2} \varphi_j L_{it}^j + \varepsilon_{it} + \omega_{it} \tag{4}$$

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.[21] Our goal is to identify the variance of the productivity shock σ_ζ^2 as well as φ_1 and φ_2. A first complication is selection effects because wages are not observed for non-participants and non-participation depends on the wage offer. Further, non-participation may depend directly on disability shocks as well as on 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.

Our selection correction is based on a reduced form rather than on our structural model, although the structural model is consistent with the reduced form. An alternative would be to include the wage risk parameters in the indirect inference estimation but this is computationally burdensome. Our reduced form model of participation is:

$$\begin{aligned} P_{it}^* &= X_{it}'\gamma + \delta_1 L_{it}^1 + \delta_2 L_{it}^2 + \theta G_{it} + \vartheta_{it} \\ &= s_{it} + \vartheta_{it} \end{aligned} \tag{5}$$

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 serve as exclusion restrictions. The unobserved "taste for work" ϑ_{it} is freely correlated with the permanent productivity component ε_{it}.

In Table 1 column (1) we report marginal effects from a probit regression for participation. Participation is monotonically decreasing in the degree of work limitations. Among the low educated, the probability of working declines by 13 percentage points at the onset of a moderate work limitation, and by 55 percentage points at the onset of a severe work limitation. Regarding our exclusion restrictions, the signs are correct: higher income from transfers and a more generous welfare system increase the opportunity cost of work, and the effects are statistically significant.[22]

[21] 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).

[22] 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.

Estimation of (5) allows us to construct an estimate of the inverse 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. Next, we estimate

$$\ln w_{it} = X_{it}'\alpha + \varphi_1 L_{it}^1 + \varphi_2 L_{it}^2 + \sigma_{\varepsilon\vartheta}\lambda(s_{it}) + v_{it} \tag{6}$$

only on the sample of workers, and with $E(v_{it}|P_{it}^* > 0, X_{it}, L_{it}) = 0$. The resulting estimates of φ_1 and φ_2, with the selection correction through the inverse Mills' ratio, should be interpreted as the estimates of the effect of work limitations on *offered* wages.

Table 1: **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)
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

The regression also includes age, age squared and dummies for marital status and year.

In columns (2) and (3) of Table 1, we report estimates of the log wage process with and without correcting for endogenous selection into work. The key coefficients are the ones on $\{L = 1\}$ and $\{L = 2\}$, which are estimates of φ_1 and φ_2, the effect of the work limitation on wages. A moderate work limitation reduces the offered wage rate by 21 percentage points, whereas a severe limitation reduces the offered wage by 40 percentage points. The selection correction to recover the offered wage from the observed wage makes a substantial difference. The effect of a severe work limitation on the observed wage is 8 percentage points less than on the offered wage: those who remain at work despite their work limitation have higher-than-average permanent income (shown by the positive sign of the Mills ratio).

4.3 Productivity Risk

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

$$g_{it} = \Delta\left(\ln w_{it} - X_{it}'\alpha - \varphi_1 L_{it}^1 - \varphi_2 L_{it}^2\right) \tag{7}$$

From estimation of α, φ_1 and φ_2 described above we can construct the "adjusted" residuals (7), and use them as they were the true adjusted error terms (MaCurdy, 1982). We can then identify the variance of productivity shocks and the variance of measurement error using the first and second moments and the autocovariances of g_{it}, as discussed fully in the online appendix. Standard errors are computed with the block bootstrap.

The results are in Table 2. The estimate of the variance of productivity shocks is similar to estimates reported elsewhere (see Meghir and Pistaferri, 2004). 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 2: **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)

4.4 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 DI policy parameters ($\pi_{L=0}^{Young}, \pi_{L=1}^{Young}, \pi_{L=2}^{Young}, \pi_{L=0}^{Old}, \pi_{L=1}^{Old}, \pi_{L=2}^{Old}$, and P^{Re}) is achieved by Indirect Inference (see Gourieroux et al, 1993).[23] Indirect inference relies on matching moments from an approximate model (known as auxiliary model) which can be estimated on both real and simulated data, rather than on moments from the correct data generating

[23] We set some parameters to realistic values: the risk aversion parameter $\gamma = 1.5$, the annualised interest rate is 1.5% and the discount rate 2.5%. A life-span is 50 years, from age 22, with the last 10 years in compulsory retirement. The quarterly arrival rate of job offers is set to 0.73. Ideally, we would identify the value of λ by using durations of unemployment by disability status. However, censoring problems, as well as noise when we stratify by education and work limitation status, lead us to take the value of λ from Low et al. (2010).

process. 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. Any bias in estimates of the auxiliary model on actual data will be mirrored by bias in estimates of the auxiliary model on simulated data, under the null that the structural model is correctly specified. However, the closer the link between the parameters of the auxiliary equations and the structural parameters, the more reliable is estimation.

In our theoretical model, individuals make three decisions: how much to consume, whether to work, and whether to apply for DI. We use auxiliary equations that reflect these choices.[24] In particular, we use: (1) the stock of recipients of DI, conditional on disability status and age; (2) the DI status of people of different age and health status; (3) a regression of log consumption on work limitation, disability insurance, participation (and interactions), controlling for a number of other covariates; and (4) labour force participation rates, conditional on disability status and age. This choice gives us 30 moments overall.

4.4.1 Moments: Disability Insurance

There are two ways in which we calculate moments involving the stock of DI recipients. First, we consider the composition of DI recipients by health status. This identifies the fraction of DI 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. For example, we use the fraction of those with a severe limitation who are in receipt of DI to help identify the fraction of the truly disabled who benefit from the insurance. This fraction is related to the parameter governing the probability of a successful application: it would be particularly informative if all $L = 2$ individuals applied and no one left the program. 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.

[24] We do not have data on DI application, and hence use moments reflecting participation in the DI program.

4.4.2 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 increases the marginal utility of consumption (e.g. through increased needs) then consumption will rise 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 non-separability 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) work limitation on consumption for individuals who are not in receipt of DI is given by the parameter α_3. This captures both the income effect and the non-separability. For individuals who are in receipt of DI, the effect of a severe disability on consumption is $(\alpha_3 + \alpha_4)$. If DI provided full insurance, $\alpha_3 + \alpha_4$ would capture the effect of the non-separability, with the parameter α_4 negating the income effect in α_3.[25] The split between α_3 and α_4 is less clear if insurance is partial; in which case $(\alpha_3 + \alpha_4)$ 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

[25] 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 income effects. However, if we could find a group of individuals who are fully insured against disability shocks, then the consumption response to the work limitation for those individuals would 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.

structural interpretation directly on the values of the α parameters. The coefficients α_1 and α_2 correspond to the effects of a moderate disability. We control for permanent income and age because we want to compare individuals facing the same level of insurance through the DI system.[26] We control for unearned income to compare individuals with the same potential for self-insurance.[27]

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-separablity between consumption and labor supply, as well as the fixed cost of work.

4.4.3 Moments: Participation over the Life-Cycle

We calculate participation rates by age and by disability status, using four 10-year age bands: 23-32, 33-42, etc. The moments are the participation rates for the three work limitation groups in each age band, giving 12 moments overall. 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. Unemployment rates among the healthy in the early life cycle are informative about the job destruction rate δ because assets are very low at young ages and so very few quit employment. The differences in participation by disability status is informative about the fixed costs of work and how these differ by work limitation status (i.e., the extent that work is more costly for disabled than for healthy workers).

4.4.4 Indirect Inference Results

In this section we present results on the moments matched by Indirect Inference and the estimates of the structural parameters generated by it. All moments are presented in the three sections of Table 3.

The first section comes from matching DI policy moments: the work limitation status of DI recipients (Panel A) and the work limitation status of DI recipients (Panel B) separately for younger (age$<$45) and older workers (age$\geq$ 45). Our model is capable of matching most

[26] 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.

[27] We need to add two caveats to our identification strategy. First, as stressed by Meyer and Mok (2008), consumption is measured at the family level, but we observe changes in disability at the individual level. To partly account for this, we use a measure of adult equivalent consumption. The second caveat is that disability insurance is only one form of insurance against disability risk (SSI and workers' compensation being others). We replicated the regression reported in section B of Table 3 using a more comprehensive measure of insurance against disability risk (comprising DI, SSI and WC) and find qualitatively similar results.

Table 3: **Matched Moments**

Section A: Disability Insurance Moments

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

Section B: The Log Consumption Regression

Variable	Data	Sims
$\{L_{it} = 1\}$	-0.121	-0.072
$\{L_{it} = 2\}$	-0.184	-0.146
$\{L_{it} = 1\}\, DI$	0.276	0.131
$\{L_{it} = 2\}\, DI$	0.486	0.260
DI	-0.278	-0.008
Employed	0.456	0.337

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

Section C: Labor Market Participation by Disability Status

Age band	No limit		Moderate		Severe	
	Data	Sims	Data	Sims	Data	Sims
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

of the moments with great accuracy. For example, it matches quite closely the proportions of "false recipients", $\text{Fr}(L=0|DI=1,t)$, as well as the proportion of workers "insured" by the DI program, $\text{Fr}(DI=1|L=2,t)$, which are the reduced form equivalents of the incentive cost/insurance benefit tradeoff. The second section of Table 3 reports the moments obtained from estimating the auxiliary log consumption equation (using imputed data, as detailed above).[28] The signs and in most cases the magnitude of the coefficients are similar. These estimates suggest that consumption falls when people become disabled and there is no insurance. However, those who are fully insured against the disability shock (those who are receiving DI) increase their consumption, consistent with the idea that consumption and poor health are Frisch complements ($\theta < 0$ in our utility specification).[29] Finally, the third section shows participation over the life cycle for people of different work limitation status. 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, especially for people with severe disability.

In Table 4 we report the Indirect Inference parameter estimates corresponding to these moments. We estimate that a moderate (severe) disability induces about a 4% (8%) loss of utility in terms of consumption. Participation induces a 32% loss.[30] The fixed costs of work per quarter rise substantially with the degree of disability. We estimate that a job is destroyed on average every 26 quarters. 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 discuss further in the next section.

---

[28] 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).

[29] In the online appendix we confirm this by conducting an event study analysis. We study the consumption behavior of individuals who at some point become severely disabled (time 0), some of whom have access to DI and some who do not. The two groups have similar consumption levels *before* the onset of disability (p-value=40%), but the group of individuals receiving DI have significantly higher consumption after the onset of disability (p-value is 2.38%). Starting from the rather large fall in the period just before the onset of disability, we find that the group of individuals with access to DI experience an *increase* in consumption over the next 6 years, while the other group experiences a further decline.

[30] 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 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).

Table 4: **Estimated Parameters**

Frictions and Preferences			Disability Insurance Program	
Paramter		Estimate	Parameter	Estimate
θ	Cost of disability	-0.039 (0.017)	$\pi^{Young}_{L=0}$	0.002 (0.00002)
η	Cost of part.	-0.32 (0.0033)	$\pi^{Old}_{L=0}$	0.009 (0.00032)
δ	Job destruction	0.049 (0.00003)	$\pi^{Young}_{L=1}$	0.103 (0.0132)
			$\pi^{Old}_{L=1}$	0.14 (0.0088)
$F_{L=0}$	Fixed cost	0.10 [$303] (0.0014)	$\pi^{Young}_{L=2}$	0.35 (0.041)
$F_{L=1}$	Fixed cost	0.31 [$942] (0.013)	$\pi^{Old}_{L=2}$	0.72 (0.0044)
$F_{L=2}$	Fixed cost	1.20 [$3646] (0.0072)	P^{Re}	0.050 (0.00038)

Fixed costs are reported as the fraction of average offered wage income at age 23 and also in 1992$ per quarter. Standard errors in parenthesis.

4.5 Implications: Flows onto and off DI

We use our model to simulate the rate of flows on and off DI by work limitation status, and we compare these to rates in the data. We did not use these in the estimation because these moments are imprecisely estimated in the data given the size of our sample. However, we reproduce in Table 5 the main flow statistics and the simulated counterparts as an indication of the performance of the model. Simulated flows on and flows off DI match the data fairly well, despite the estimation being based only on the stock variables.

4.6 Implications: 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 unconditional Award rate: $\Pr(DI = 1|DI^{App} = 1)$. We simulate 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% average success rate.[31] Our estimate contrasts quite well also with the reduced form estimates (of 0.45) obtained by Bound and Burkhauser (1999) and others using data on individual DI application and DI receipt from the HRS.

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

Table 5: **Flows onto and off Disability**

	Data	Simulations
Flows onto DI		
$\text{Fr}(DI_t = 1 \mid DI_{t-1} = 0, L_t = 2, t < 45)$	0.12	0.12
$\text{Fr}(DI_t = 1 \mid DI_{t-1} = 0, L_t = 2, t \geq 45)$	0.19	0.29
$\text{Fr}(DI_t = 1 \mid DI_{t-1} = 0, L_t = 1, t < 45)$	0.0055	0.016
$\text{Fr}(DI_t = 1 \mid DI_{t-1} = 0, L_t = 1, t \geq 45)$	0.033	0.023
Flows off DI		
$\text{Fr}(DI_t = 0 \mid DI_{t-1} = 1, L_t = 2, t < 45)$	0.109	0.109
$\text{Fr}(DI_t = 0 \mid DI_{t-1} = 1, L_t = 2, t \geq 45)$	0.079	0.049

Given that the true disability status of an applicant is private information, SSA evaluators are likely to commit two types of errors: Admitting into the DI program undeserved applicants and rejecting those who are truly disabled. Our estimates show how large are the probabilities associated with these errors. Consider first the extent of false positives (the proportion of healthy applicants who receive DI). From Table 4, these type II errors have probabilities ranging from 0.2% (young non disabled) to 14% (older workers with a moderate disability). Similarly, we can use our model to estimate the Award Error: the fraction of successful applicants to DI who are not severely disabled, given by $Pr(L = \{0, 1\} | DI = 1, DI^{App} = 1) = 0.10$. 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 4, our estimate is that the type I errors are 65% for the younger and 28% for the older workers. The fraction of rejected applicants 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 0.52-0.58, and Nagi (1969), 0.48. These comparisons confirm that our structural model is capable of replicating reduced form estimates obtained using direct information on the application and award process. We estimate the award process to be 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.[32] 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 calculate a probability of re-assessment of 7%.

5 Reform of the DI Process

The most important use of our model and structural estimates is the ability to analyse the effects on welfare and behaviour of changing the main parameters of the DI program. We consider four changes: first, changing the generosity of disability payments; second, making the program "stricter" by increasing the threshold that needs to be met in order to qualify for benefits; third, changing the reassessment rate of disability recipients; and finally, changing the generosity of the food stamp program. For each scenario, we study the implications for welfare, for the efficiency of the DI process and for behaviour more generally. We calculate the welfare implications by measuring the willingness to pay for the new policy through a proportional reduction in consumption, π, at all ages which makes the individual indifferent between the status quo and the policy change considered.[33] In all the experiments below the impact on the government budget is neutralised by adjusting the proportional wage tax iteratively (see equation (3) in the online appendix).

5.1 Generosity of DI Payments

We consider the effects of revenue-neutral, proportional changes in DI generosity, with the proportional changes ranging from a cut to 30% of its current value to a 50% increase. Figure 2 shows the effects of these changes, and table 6 reports elasticities evaluated at the baseline.

Column 1 of Table 6 reports the elasticity of applications with respect to the generosity of disability insurance. The elasticity for all individuals is 0.69, which is a central estimate from those available in the literature using micro data (see Bound and Burkhauser, 1999). What is striking, however, is how this elasticity differs when we consider moderate and severely

[32] 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.

[33] This is obtained by calculating expected utility at the start of the life-cycle before the resolution of any uncertainty ("behind the veil of ignorance").

disabled individuals on their own: the moderately disabled are very elastic in their response to generosity, whereas the severely disabled have essentially no response. The slight negative elasticity for the severely disabled arises because individuals choose to apply when they are only moderately disabled, and so the pool of severely disabled workers who have not already applied for DI is a selected group of higher productivity individuals.

The second column shows the elasticity of the employment rate with respect to benefit generosity. This elasticity is low, with very little effect of generosity on overall labour supply, similar to the results of Bound (1991). There is also a differential effect on the moderately and severely disabled. The dynamics here are very important: participation is lower among the severely disabled even though applications have fallen because a number of severely disabled who are not working, stopped work and applied when they were only moderately disabled. Similarly, the small positive effect on participation among the moderately disabled arises because many are turned down for DI and have to go back to work.

Table 6: **Implied Elasticities**

	$Fr\left(Apply_t \middle| L_t, DI_{t-1}=0\right)$ Elasticity	$Fr\left(Work_t \middle| L_t\right)$ Elasticity
All Individuals	0.69	-0.025
Moderate disability: $L=1$	2.10	0.025
Severe disability: $L=2$	-0.091	-0.35

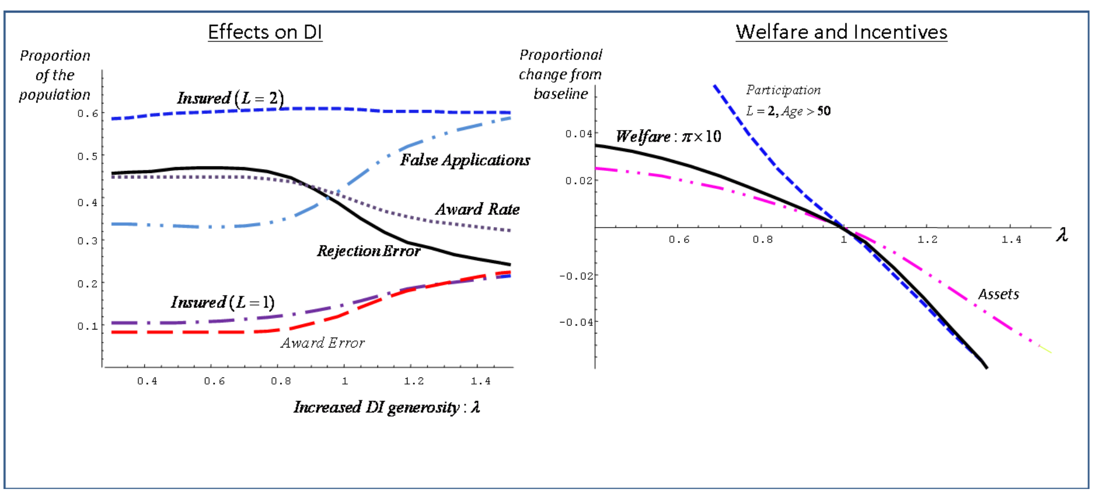

Figure 2: Changing DI Generosity

Figure 2 provides more detail on the behavioral effects. Increasing the generosity of DI payments increases sharply the fraction of applicants who are not severely disabled (the "False Applications" line on the left-hand side). This in turn leads to an increase in the award error and in the fraction of the moderately disabled who are receiving insurance (the "Insured ($L = 1$)" line shows this fraction for those 45 and over). The fall in the rejection error arises mechanically: greater numbers of false applicants mean the fraction of the rejections who are severely disabled falls. What is striking is that there is very little change in the fraction of the severely disabled who receive insurance (the line "Insured ($L = 2$)"), and this is because applications for DI from this group are insensitive to the generosity of DI as discussed earlier.

Given these effects, the welfare implications of changing generosity shown in the right hand graph of Figure 2 are not surprising. For all statistics considered on the right-hand side, the y-axis measures the proportional change relative to the baseline.[34] Increases in DI generosity funded by a wage tax reduce welfare, and a 10% increase in generosity implies a welfare loss of 0.13% of consumption. The broader incentive effects of changing generosity vary by work limitation status: for the severely work limited, greater generosity has the direct effect of encouraging applications for DI and individuals move out of the labour force. The greater generosity also reduces asset accumulation, and this has the indirect effect of increasing participation among those who are rejected, particularly among the moderately work limited.[35]

5.2 Strictness of DI Admissions

Increases in strictness of DI in 1980 led to sharp declines in inflows onto DI and significant numbers of recipients being removed, although the criteria was relaxed again in 1984. The issue is whether the benefit that greater strictness has of reducing incentives for false applications outweighs the worsening insurance. To tackle this issue, we need first to define a

[34] We show participation rates only for those over 50 because the effects on participation at earlier ages are qualitatively similar. The line "Assets" shows how the maximum average asset holding over the life-cycle varies.

[35] Our results differ from Meyer and Mok (2008), who apply a variant of the benefit optimality formula derived by Chetty (2008) to conclude that the current level of DI benefits is lower than the optimal level. What explains the discrepancy? The formula requires estimates of risk aversion and prudence, of the fall of consumption on disability, and an estimate of the elasticity of DI application with respect to DI benefit generosity. Hence, there are three reasons why our results differ from theirs. First, the formula imposes a common elasticity of DI application to benefits without distinguishing between the disability status of applicants. But as shown in Table 6, elasticities are dramatically different when conditioning on work limitation status. Second, with non-separabilities, the fall of consumption upon disability understates the value of insurance. Finally, we assume a more moderate degree of risk aversion (1.5 vs. 3).

measure of strictness of the program.

Suppose that Social Security DI evaluators decide whether to award DI as a function of a noisy signal about the severity of the applicant's disability status, which has some distribution:

$$S_{it} \sim f(L, t)$$

The properties of the distribution of the signal S vary by age (for simplicity, for two age groups defined by age<45 and age≥45), and by work limitation status L. The Social Security DI evaluators make an award if $S_{it} > \overline{S}$. The parameter $\overline{S}$ 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 assume that S lies between 0 and 1 and has a Beta distribution, $\beta(a_{L,t}, b_{L,t})$, whose parameters a and b vary with age and work limitation status. The values of $a_{L,t}$ and $b_{L,t}$ and of $\overline{S}$ are pinned down by the six structural probabilities (π_L^t) estimated above:[36]

$$\begin{aligned} 1 - \pi_L^t &= \Pr(\text{Rejection} | t, L, Apply) \\ &= CDF(\beta(a_{L,t}, b_{L,t})) \end{aligned}$$

Figure 3 illustrates the resulting distributions of S for those over 45 by work limitation status, and illustrates some of the errors under the estimated DI program. The area on the left of $\overline{S}$ under the dashed light grey curve (labeled $f(S|L = 2, t \geq 45)$ measures the probability of rejecting a deserving DI applicant. The area on the right of $\overline{S}$ under the solid grey curve (labeled $f(S|L = 1, t \geq 45)$) measures the probability of accepting into the DI program a DI applicant with only a moderate disability. Increasing the strictness of the test (increasing $\overline{S}$) reduces the probability of false positives (reduces the extent of the incentive problem), but increases the probability of false negatives (reduces the extent of insurance provided by the program). It also can have substantial effects on who applies. A policy of changing $\overline{S}$ therefore has both benefits and costs, trading off incentives against

[36] We normalise the mean of the signal, S, for the old who are severely disabled and the mean of S for the young who are not at all disabled to being 0.6 units apart, and we impose that the parameter b is identical across age and work limitation status. These normalisations, alongside the use of the Beta distribution, impose a particular distribution on the signals which we do not have the data to test. We considered alternative assumptions, such as a normal distribution with age and disability shifting the mean of the signal, and our results are qualitatively similar. The intuitive advantage of the Beta distribution is that the precision of the signal increases as true disability status worsens.

insurance, and we use our model to determine which dominates when the strictness of the test changes.[37]

Figure 3: The Distrbution of S for the Older Worker by Work Limitation Status

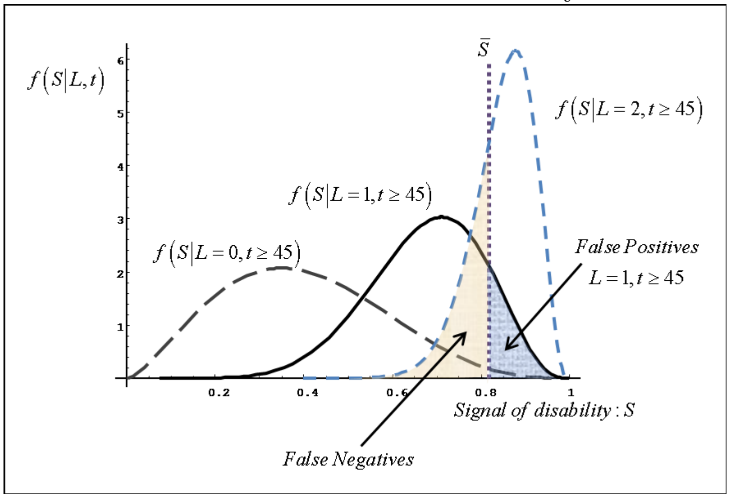

Figure 4 reports the results of this experiment. The left-hand graph shows the implications for the DI program, the right-hand graph shows implications for welfare, participation and asset accumulation. Increasing $\bar{S}$ from 0.65 to 0.95 reduces the probability of acceptance for the severely disabled over 45 from close to 100% to less than 10%. This has a direct effect of increasing the rejection error as $L = 2$ individuals are more likely to be rejected. Furthermore, the increase in $\bar{S}$ reduces the proportion of applicants from those with no or only a moderate disability. This is shown in the downward sloping broken line (labelled "False Applications"), and this implies a fall in the actual number of healthy who are rejected. Corresponding to this fall in healthy applicants and lower rate of acceptance, there is a clear decline in the fraction of awards being made to the healthy or moderately disabled (the Award Error). Conditional on the composition of applicants, increased strictness means fewer applicants are made awards, but the composition of applicants also changes, with fewer false applicants, and this means that the fraction of awards made does not decline monotonically as strictness increases (the Award Rate). The cost of increasing strictness is seen in the decline, as $\bar{S}$ increases, of the fraction of the severely work limited who are insured (the line labeled "Insured ($L = 2$)").

The right hand graph shows the incentive effects of the alternative $\bar{S}$, as well as the

[37] An alternative policy might be to reduce the noise involved in the evaluation of the signal. We do not evaluate such a policy. In theory, we could take the cost of extra SSA evaluations as being the same as the cost of a review. However, the difficulty is estimating the effect of evaluations on reducing the noise.

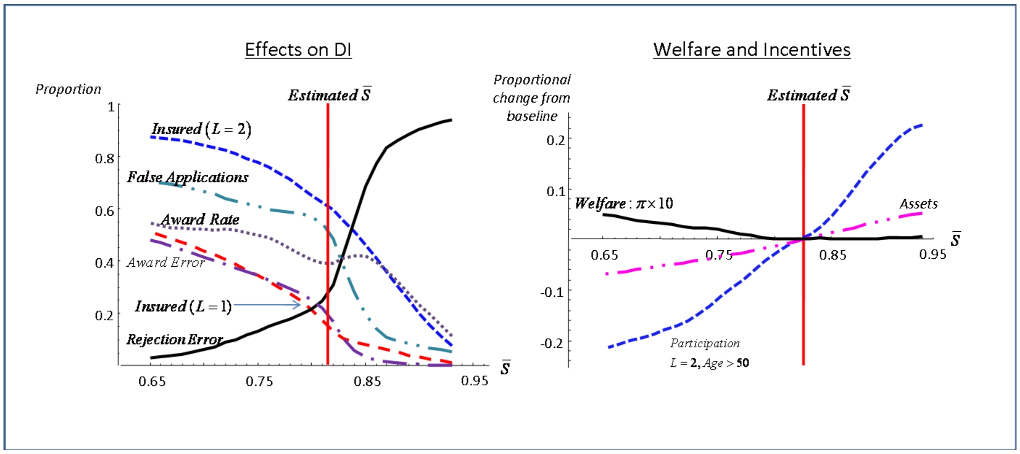

Figure 4: Changing strictness

willingness to pay. There is a direct effect of greater strictness leading to greater participation in the labor force as more people are rejected or discouraged from applying. This is particularly apparent for the severely work limited. For the moderately work limited, there is an offsetting effect: as strictness increases, individuals expect to have to self-insure and so accumulate more assets. These assets reduce participation rates among those who are rejected by DI, offsetting the direct effect of greater participation. The effects on participation for those who are not work limited at all are negligible.

The willingness to pay increases as $\bar{S}$ decreases from its estimated value: the gain in improved insurance from making the program less strict dominates the loss associated with increased numbers of false applicants and a greater award error. The magnitude of the gain in terms of consumption equivalent arising from reducing strictness from its estimated value to $\bar{S} = 0.65$ is about 0.005 (0.5%). This gain is the net gain of two offsetting effects: there is a benefit of increased insurance against disability which individuals are willing to pay for, but this is partly offset by a loss arising from output being lower as individuals work less. Part of the benefit of the relaxed strictness arises from the moderately disabled and the severely-disabled young being offered better insurance. The key to this conclusion of reduced strictness being welfare increasing is, however, the low acceptance rate of severely disabled individuals onto DI in the baseline.[38] The subgroup of young severely-disabled individuals are particularly ill-equipped to insure against disability risk because these individuals face high rejection rates when applying for DI and yet have not had time to accumulate enough

[38] We have considered various alternative specifications for how the distribution of the noisey signal varies with work limitation status and this conclusion remains. See also Denk and Michau (2010) for a similar result obtained using a dynamic mechanism design approach to the insurance-incentive tradeoff.

assets to self-insure.

5.3 Reassessment of DI Recipients

In Figure 5, we consider changing the reassessment rate. Given our estimate of the cost per reassessment,[39] this has a direct impact on the budget, as well as the effect induced by changes in the number of recipients and in labour supply. These effects are again neutralised through adjusting the wage tax. We assume that the probabilities of success, conditional on work limitation status and age, are the same at reassessment as at initial application.

The left-hand graph shows that an increase in the reassessment rate discourages false applications by those who are not severely disabled: an increase in the reassessment rate from a 0.02 probability per quarter to a 0.08 probability, leads to a decline in the proportion of applications which are false from 54% to 30%: the threat of reassessment has a marked discouragement effect, and leads to a decline in the award error, and a decline in the fraction of the non-work limited who receive insurance. The cost of this is the reduced coverage for the severely disabled: reassessment causes some severely disabled to be removed from DI and this directly reduces coverage, as well as discouraging applications, as the frequency of reassessment increases.

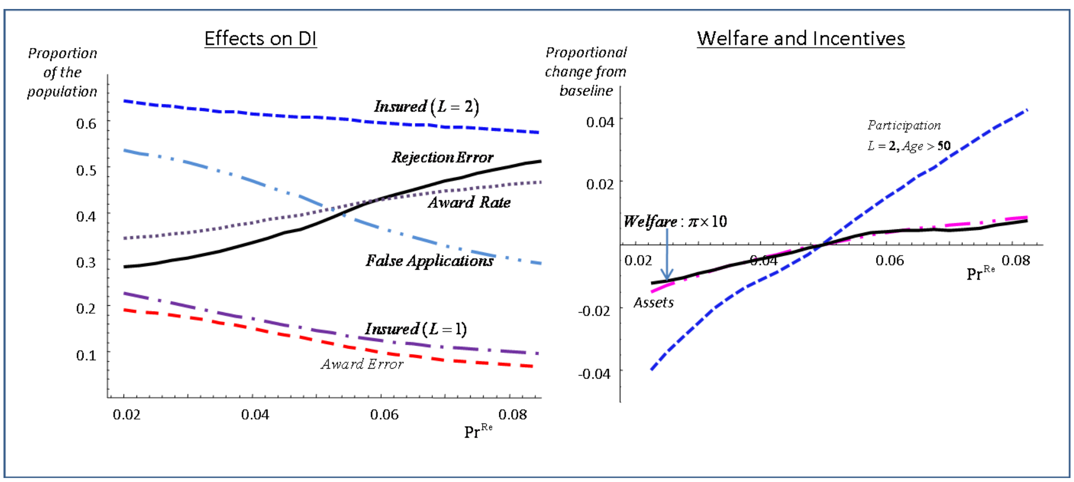

Figure 5: Changing Reassessment Rates

Despite this cost, increasing the reassessment rate increases welfare, albeit modestly, with the consumption equivalent of increasing reassessment from the baseline of 0.05 to 0.06 being

[39] For the period 2004-2008, the SSA spent \$3.985 billion to conduct 8.513 million "continuing disability reviews". This means a review costs on average \$468, and we deflate this back to 1992 prices and include this price in the government's budget constraint.

0.043%. Increased reassessment increases labor force participation among the severely work limited, who are discouraged from applying or removed from the DI rolls. This also leads to greater asset accumulation.

5.4 Generosity of The Food Stamp Program

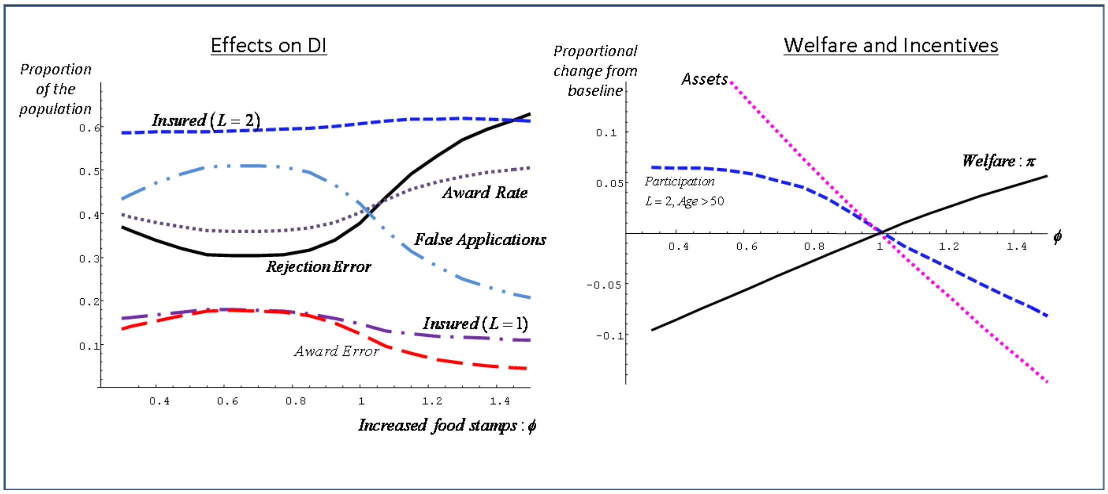

Figure 6: Changing the Generosity of Food Stamps

Figure 6 shows the effects of changing the generosity of food stamps. Increases in food stamps have a non-monotonic effect on the number of false applications: when food stamps are very low, the waiting period for a decision about a DI application is costly for those of low productivity and they do not apply. Increasing food stamps (a "consumption floor") mitigates this cost, and leads to greater numbers of false applicants. After a point, however, food stamps provide sufficiently generous support that false applications for DI fall. This effect translates into a decline in the fraction of those not severely disabled who are in receipt of DI (the "Insured ($L = 1$)") and a decline in the award error as food stamps become sufficiently generous. By contrast, the fraction of the severely disabled who receive DI increases as food stamps become more generous: this highlights the beneficial effect of food stamps making it less costly for the severely disabled to remain out of work and to apply for DI. In addition, more generous food stamps provide direct insurance against low (permanent) productivity with no risk of rejection. Together, these effects imply substantial welfare increases as the generosity of food stamps increases. A 10% increase in generosity implies a welfare gain of 1.4% of consumption. This is despite the fall in participation and the fall in saving that greater generosity induces for all types. It is important to stress that this movement onto food stamps is funded by a change in the tax rate and so, although the saving on DI may appear a false saving because of the greater spending on the food stamp programme, our calculations are that this is welfare increasing despite the tax rise required. What this simulation highlights is the value of food stamps in providing long term support for those whose productivity is too low to be able to work for a reasonable wage. Part of the reason for this result is that the food stamps program is less distortionary than DI because it does not require people to disengage from the labor force and to stop working altogether.

6 Conclusions

In this paper, we provide a life-cycle framework for estimating the extent of work-limiting health risk that individuals face and for analysing the effectiveness of government disability insurance against that risk. Work limitations have substantial effects on wages, with wages falling by 40% for the severely work limited. Government insurance against these shocks is incomplete: There are substantial false rejections. We estimate that 26% of the older workers with a severe work limitation who apply for benefits are rejected. This is alongside other negative effects, with some workers discouraged from applying because of the uncertainty surrounding the application process. Similarly, there are large rates of false acceptances, with between 10 and 14% of applications from those with a moderate work limitation being accepted.

We use the model to simulate various policy changes aimed at improving the insurance and mitigating the incentive costs of DI. The simulations show that the number of moderately disabled individuals receiving DI is particularly sensitive to the policy parameters, whereas the number of severely disabled is less sensitive. Thus, reducing DI generosity leads to a fall off in false applications and mis-directed insurance, without reducing applications from the severely disabled who are essentially inelastic with respect to benefit generosity. Of course, the severely disabled will then receive less insurance, but this reduced generosity increases welfare, at least from an ex-ante perspective. On the other hand, increasing the strictness of the DI screening process leads to a decline in welfare because the existing program already suffers from turning down large numbers of severely disabled, especially among the young. Increasing the generosity of Food Stamps leads to a fall off in false applications for DI and mis-directed insurance, leading to better targeting of DI and a welfare improvement despite the extra cost of Food Stamps. More frequent reassessments of recipients directly reduces the number of recipients who are not severely work limited, but equally importantly more frequent reassessments substantially reduces the proportion of false applicants, leading to welfare gains. In summary, welfare increases if the threshold for acceptance is lower, disability payments are lower, reassessment more frequent and food stamp payments more generous. These conclusions arose because these reforms lead to a separation of the severely work limited from the moderately limited for whom work is a realistic option. This highlights the need to have disability classified into more than just a "yes" or "no" state, and raises the question of whether allowing for partial disability and partial DI payments (as in the Netherlands, for example) may be a way to reduce the "moral hazard" problem of DI.

One limitation of the conclusions in the previous paragraph is the clear non-linearities in behaviour apparent from the simulations in section 5. This highlights the value of having careful structural models of behaviour in analysing disability shocks and the DI process.

One of the implications of our simulations is that changes to the DI process can have sizable effects on asset accumulation, both by changing the need for self-insurance and by changing the amount of time that individuals spend out of the labour force. Related to this, Golosov and Tsyvinski (2006) propose that an asset-test should be introduced to the DI award process to identify those applicants who accumulated assets explicitly to smooth consumption while faslely claiming DI. We could in principle explore an asset test in our framework and whether an asset test discourages applicants among the moderately or severely disabled, the difficulty is that assets in our framework are fully fungible and serve multiple purposes, including retirement saving, general consumption smoothing as well as self-insurance. An asset test for DI applicants would therefore have the unfortunate side effect of reducing retirement saving.

In terms of further extensions, our model of the disability insurance process is incomplete: Benitez-Silva et al. (2004) have emphasized the importance of the appeal process, whereas we have allowed the social security administration to make just one decision. In the context of capturing behaviour over the life-cycle this may be less problematic, but it means we cannot examine one dimension of reform, namely the strictness and length of the appeal judgement relative to the initial judgement. A second restriction is in terms of the stochastic process for work limitations, which we take to be exogenous. The probability of receiving a negative shock to the ability to work is likely to be partly under individuals' control, through occupation choice and other decisions on the job. These decisions will be affected by the properties of the disability insurance scheme. Finally, we have ignored the health insurance component of the program (although our fixed cost for work process could be partly re-labeled to capture health spending differences by health and employment status). This means we underestimate the insurance value provided by the program.

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] Attanasio, O., and G. Weber (1995), "Is Consumption Growth Consistent with Intertemporal Optimization? Evidence from the Consumer Expenditure Survey", Journal of Political Economy, 103(6), 1121-57.

[4] Benitez-Silva, H., M. Buchinsky, H. M. Chan, S. Cheidvasser, and J. Rust (2004), "How Large Is the Bias in Self-Reported Disability?", Journal of Applied Econometrics, Vol. 19 (6), 649-670.

[5] Benitez-Silva, H., M. Buchinsky, H. M. Chan, S. Cheidvasser, and J. Rust (1999), "An Empirical Analysis of the Social Security Disability Application, Appeal and Award Process", Labour Economics 6 147-178.

[6] Benitez-Silva, H., M. Buchinsky, and J. Rust (2006a), "How Large are the Classification Errors in the Social Security Disability Award Process?", NBER Working Paper 10219.

[7] Benitez-Silva, H., M. Buchinsky, and J. Rust (2006b), "Induced Entry Effects of a $1 for $2 Offset in SSDI Benefits", unpublished manuscript.

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

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

[11] Bound, J. and R. V. Burkhauser (1999), "Economic Analysis of Transfer Programs Targeted on People with Disabilities", 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] Bound, J., Stinebrickner, T and T.Waidmann (2010) “Health, economic resources and the work decisions of older men” Journal of Econometrics 156: 106-129

[14] Bound, J., and A. Krueger (1994) “The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?,” Journal of Labor Economics, 9, 1-24.

[15] Burkhauser, R.V. and M. 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.

[16] Chen, S. and W. van der Klaauw (2008) “The Effect of Disability Insurance on Labor Supply of Older Individuals in the 1990s", Journal of Econometrics, Vol. 142(2), p.757-784.

[17] Chetty, R. (2008), "Moral Hazard vs. Liquidity and Optimal Unemployment Insurance", Journal of Political Economy, 116(2): 173-234.

[18] Daly, M. C. and R.V. Burkhauser (2003) “The Supplemental Security Income Program” in Means-Tested Programs in the United States,” edited by Robert Moffitt. National Bureau of Economic Research and University of Chicago Press: Chicago, IL.

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

[20] DeNardi, M., French, E. and J. Jones (2010), "Why Do the Elderly Save? The Role of Medical Expenses" Journal of Political Economy, forthcoming.

[21] Denk, O. and J.-B. Michau (2010), "Optimal Social Security with Imperfect Tagging", mimeo.

[22] Diamond, P. and E. Sheshinski (1995), “Economic aspects of optimal disability benefits”, Journal of Public Economics 57 (1): 1-23.

[23] Duggan, M. and S.A. Imberman (2009), "Why Are the Disability Rolls Skyrocketing? The Contribution of Population Characteristics, Economic Conditions, and Program Generosity", in *Health at Older Ages: The Causes and Consequences of Declining Disability among the Elderly*, NBER.

[24] 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.

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

[26] Golosov, M. and A. Tsyvinski (2006), “Designing Optimal Disability Insurance: A Case for Asset Testing”, Journal of Political Economy 114, 257-279.

[27] Gourieroux, C., A. Monfort, and E. Renault (1993), "Indirect Inference", Journal of Applied Econometrics, Vol. 8, Supplement: Special Issue on Econometric Inference Using Simulation Techniques, pp. S85-S118.

[28] Hoynes, H.W. and R. 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).

[29] Kreider, B. (1999), "Latent Work Disability and Reporting Bias," Journal of Human Resources, 734-769.

[30] Lillard, L.A., and Y. Weiss (1997), "Uncertain Health and Survival: Effects on End-of-Life Consumption." Journal of Business and Economic Statistics, 15(2): 254–68.

[31] Low, H., C. Meghir and L. Pistaferri (2010), "Wage risk and employment risk over the life cycle", American Economic Review, forthcoming.

[32] MaCurdy, T. (1982) "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis" Journal of Econometrics 18(1): 83-114.

[33] Meghir, C. and L. Pistaferri (2004) "Income variance dynamics and heterogeneity " Econometrica.

[34] Meyer, B.D. and W. K.C. Mok (2007) "Disability, earnings, income and consumption" University of Chicago, mimeo

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

[36] Parsons, D.O. (1980), "The Decline in Male Labor Force Participation", The Journal of Political Economy 88, No. 1: 117-134

[37] Rust, J., M. Buchinsky, and H. Benitez-Silva (2002), "Dynamic Models of Retirement and Disability", Working Paper.

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

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

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

[41] Waidman, T., J. Bound and A. 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.