Disability Risk, Disability Insurance and Life Cycle Behavior

IFS Working Paper W10/11

Hamish Low
Luigo Pistaferri

Disability Risk, Disability Insurance and Life Cycle Behavior*

Hamish Low
University of Cambridge

Luigi Pistaferri
Stanford University

April 14, 2010

Abstract

This paper provides a life-cycle framework for weighing up the insurance value of disability benefits agains the incentive cost. Within this framework, we estimate the life-cycle risks that individuals face in the US, as well as the parameters governing the disability insurance program, using indirect inference and longitudinal data on consumption, disability status, disability insurance receipt, and wages. We use our estimates to characterize the effectiveness of the disability insurance program and to consider the effect on welfare and behaviour of policy reform. High levels of false rejections associated with the screening process imply welfare increases as the program becomes less strict, despite the worsening incentives that this implies. Incentives for false applications are reduced by reducing generosity and increasing reassessments and these improve welfare, despite the worse insurance implied.

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, 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 Autor and Duggan, 2006). The key questions in thinking about the size and growth of the program are whether program claimants are genuinely unable to work, and how valuable is the insurance provided. These questions underlie the concerns that the greater use of DI explains the recent decline in labor market participation of men and that the DI program is being used as a gateway for early retirement, rather than providing insurance against health shocks that prevent work. To assess these concerns and to evaluate the costs and benefits of changing the DI program to try to reduce disincentives to work, we need a realistic modeling of the risks that individuals face over their life-cycle to their health and ability to work, and a theoretical framework that captures both the insurance benefit of DI as well as the incentive effects on individual behavior. The broad aim of this paper is to provide this framework and to use it to evaluate the DI program quantitatively in an explicit life-cycle setting.

More specifically, our paper makes three contributions. First, we propose a theoretical framework that allows us to study in a life cycle setting 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 key 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 status, 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 so the

[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.

use of the PSID is important for studying this behavior and behavior across the whole life-cycle, rather than using the HRS which is restricted to those aged over 50. Our estimates highlight 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 false applications, labour supply and savings to changes in the details of the DI program, and what are the welfare implications. 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) stringency of the screening process, (b) re-assessment rate, (c) replacement rate, (d) 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 ex-ante welfare increase when the program becomes less strict, despite the increase in false applications that this implies. 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 false applicants at little cost in terms of reduced coverage for those in need.

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 small 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.

Suggested reforms to reduce these inefficiencies either recommend preventing false applications in the first place, or providing incentives or mechanisms to make false claimants leave the programme. At the heart of this different focus is the question of whether those receiving DI unjustly

were healthy (and hence could have worked) when they entered the programme, or whether their health improved while on DI but they chose not to leave. 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 Gastwirth, 1972; 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 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 programme. DeNardi, French and Jones (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, Banks and Gaudecker (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), 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.[2] 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. Our framework includes an explicit measure of disability risk, 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. Finally, the opportunity cost of applying for DI depends on whether there are programs to finance consumption during the period it takes for an application to be processed, 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. The key sections of the paper are Sections 5 and 6. Section 5 presents the estimates of the structural parameters and discusses the efficiency of the existing DI system. Section 6 carries out counter-factual policy experiments, reporting the effects on behavior and welfare of potential reforms of DI. Section 7 concludes.

2 Life-Cycle Model

2.1 Individual Problem

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

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

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

where β is the discount factor, E_t the expectations operator conditional on information available in period t (a period being a quarter of a year), P a discrete $\{0,1\}$ labor supply participation 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 show 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

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

where A are beginning of period assets, R is the interest factor, w the hourly wage rate, h a fixed number of hours (corresponding to 500 hours per quarter), τ_w a proportional tax rate that is used to finance social insurance programs, F the fixed cost of work that depends on disability status,[3] 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^{UI}, E^{DI}, and E^W are recipiency $\{0,1\}$ indicators for unemployment insurance, disability insurance, and the means-tested welfare program, respectively.[4]

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

We use a utility function of the form

[3]The fact that disabled individuals face direct costs of work is explicitly recognized 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).

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

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

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.[5]

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, Meghir and Pistaferri (2010). The innovations in our set-up are to consider the risk that arises from disability shocks that cause a work limitation and the explicit modelling of disability insurance.

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 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.[6] 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 is on the low educated. To simplify notation, we omit subscripts defining the skill group of interest.

[5] 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.

[6] See the Web Appendix for more details. This is available at http://www.stanford.edu/~pista/papers/WA_LP.pdf.

2.2 The Wage Process and Labour Market Frictions

We model the wage process for individual i as being subject to general productivity shocks and shocks to work limitation status, as well as the contribution of observable characteristics X_{it}:

$$\ln w_{it} = X'_{it}\alpha + \varphi_1 L^1_{it} + \varphi_2 L^2_{it} + \varepsilon_{it} \tag{1}$$

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

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

Individuals work limitation status, L_{it}, evolves according to a three state first-order Markov process. 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.[7] 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.

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

This wage and employment environment implies a number of sources of risk, from individual productivity, work limitation shocks, and 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.

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

2.3 Social Insurance

2.3.1 The SSDI Program

The Social Security Disability Insurance program (SSDI) is an insurance program for covered workers, their spouses, and dependents that pays benefits related to average past earnings. The purpose of the program is to provide insurance against health shocks that impairs 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.[8]

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 at least 12 months.*"[9]

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). If such disability is a listed impairment, the individual is awarded benefits without further review.[10] 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

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

[9] 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 was 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.

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

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 one period prior to becoming unemployed and making the application. Third, individuals must have been unemployed for at least one quarter before applying. Successful applicants begin receiving benefits in that second quarter. Unsuccessful individuals must wait a further quarter before being able to return to work, but there is no direct monetary cost of applying for DI. Finally, we assume that the probability of success depends on the true work limitation status, age, and education:

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

The medical requirement in the SSDI program imposes a severity and persistence requirement on the work limitation. In our model, the expected persistence of the work limitation is captured by the Markov process assumption and is age dependent. This age dependence is the reason why we make the probability of a successful application for DI dependent on age.[11] The survey question we use to identify the work limitation (described fully below) asks individuals about *work*-related limitations rather than medical conditions or health status more generally. 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. The probability of being reassessed is 0 for the first year, then is given by P^{Re}, which is independent of L and age. If an individual is not successful on application or if an individual is rejected on reassessment, the individual has to remain unemployed until the next quarter before taking up a job. Individuals can only re-apply in a subsequent unemployment spell.

SSDI benefits are calculated in essentially the same fashion as Social Security retirement benefits. Beneficiaries receive indexed monthly payments equal to their Primary Insurance Amount (PIA), which is based on taxable earnings averaged over the number of years worked (known as AIME). Caps on the amount that individuals pay into the DI system as well as the nature of the formula determining benefits (see equation 2 below) make the system progressive. Because of the

[11] 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).

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. However, benefits are independent of the extent of the work limitation.

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

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

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

To understand our characterization of the application process and the trade-off between genuine applicants and non-genuine applicants, consider the following example. Assume that the government receives a noisy signal S_{it} about the true disability status of a DI applicant (independent of non-health related productivity ε_{it}), and that its decision rule is to award benefits to applicants whose signal exceed a certain stringency threshold, $S_{it} > \overline{S}$. 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 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 similar 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.

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

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.

2.3.2 Unemployment Insurance

We assume that unemployment benefits are paid only for the quarter immediately following job destruction. Unemployment insurance is paid only to people who have worked in the previous period, and only to those who had their job destroyed (job quitters are ineligible for UI payments, and we assume this can be perfectly monitored). We assume $B_{it} = b \times w_{it-1}\overline{h}$, subject to a cap, and we set the replacement ratio $b = 75\%$. This replacement ratio is set at this high value because the payment that is made is intended to be of a similar magnitude to the maximum available to someone becoming unemployed. This simplifying assumption means that, since the period of choice is one quarter, unemployment benefit is like a lump-sum payment to those who exogenously lose their job and so does not distort the choice about whether or not to accept a new job offer. Similarly, there is no insurance against the possibility of not receiving a job offer after job loss.

2.3.3 Universal Means-Tested Program

The universal means-tested program is an anti-poverty program providing a floor to income for all individuals, similar to the actual food stamps program but with three important differences. First, means-testing is on household income rather than on income and assets; second, the program provides a cash benefit rather than a benefit in kind; and third, we assume there is 100% take-up.[13] Gross income is given by

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

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

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

value of the program is then given by

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

The maximum value of the payment, $\overline{T}$, is set assuming a household with two adults and two children. The term $\underline{y}$ is the poverty line and so only people with net earnings below the poverty line are eligible.

As we discuss below, this means-tested program interacts in complex ways with disability insurance: the Food Stamps program provides a consumption floor during application for DI, and an alternative mechanism for income support for those of low productivity.

2.3.4 Supplemental Security Income (SSI)

Individuals who are deemed to be disabled according to the rules of the DI program and who have income (comprehensive of DI benefits but excluding the value of food stamps) below the threshold that would make them eligible for food stamps receive also supplemental security income (SSI). The SSI program in the US is designed to help aged and disabled people who have little or no 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.[14] We assume that SSI generosity is identical to the means-tested program.

2.3.5 Taxation on Earnings

We hold the government budget deficit, $\bar{D}$, constant when varying parameters of the social insurance programmes. This is achieved by adjusting the proportional payroll tax, τ_w, such that the present discounted value of net revenue flows is constant:

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{R^t}\left[\left(B_{it}E_{it}^{UI}\left(1-E_{it}^{DI}\right)+D_{it}E_{it}^{DI}\right)\left(1-P_{it}\right)+E_{it}^{T}T_{it}+q\left\{\text{Re}_{it}=1\right\}\right]=\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{R^t}\tau_w w_{it}hP_{it}+\bar{D} \tag{5}$$

where q is the cost of undertaking a reassessment of an individual on disability, and Re_{it} is an indicator of whether such a reassessment has been undertaken.[15] This is done iteratively because labor supply and DI application decisions change as a consequence of changes in government policy.

[14] 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 (see note 14).

[15] 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.

2.4 Solution

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

We start by constructing the value functions for the individual when employed and when out of work. When employed, the state variables are $\{A_{it}, \varepsilon_{it}, L_{it}\}$, corresponding to current assets, individual productivity and work limitation status. We denote the value function when employed as V^e. When unemployed, there are three alternative discrete states the individual can be: unemployed and not applying for disability (denoted V^n); unemployed and applying for disability (V^{App}); or unemployed and already receiving disability insurance (V^{Succ}). We describe here the specification of the value function when employed and leave the discussion of the other value functions to the Web Appendix. The value function if working can be written as:

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

where DI_{it+1}^{Elig} is an indicator for whether the individual is eligible to apply for DI. Our model has discrete state variables for: Wage productivity, Work limitation status, Participation, Eligibility to apply for DI (if not working), and Length of time on DI (over 1 year or less than 1 year). The only continuous state variable is assets. In the value functions above, the choice of participation status is determined by the maximum of the conditional value functions.[16]

2.5 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

[16]Value functions are increasing in assets A_t but they are not necessarily concave, even if we condition on labor market status in t. The non-concavity arises because of changes in labor market status in future periods: the slope of the value function is given by the marginal utility of consumption, but this is not monotonic in the asset stock because consumption can decline as assets increase and expected labor market status in future periods changes. This problem is also discussed in Lentz and Tranaes (2005) and in Low et al. (2010).

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_η^2. The second set is labor market frictions: The job destruction rate δ, the arrival rate of job offers when unemployed λ, and the fixed cost of work $F(L)$. The third set of parameters characterize the DI policy parameters: The probability of success 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 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 Web Appendix). The match is good also in the time series. In the population, the proportion of people on DI has increased from 2.4% to 4.3% between 1985 and 2005. In the PSID the increase between 1985 and 2005 is from 2.4% to 4.5%.

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

[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 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.

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 for each individual.

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

The validity of these self-reports is somewhat controversial for two reasons: first, individuals may over-estimate their work limitation in order to justify their disability payments or their non-participation in the labour force. Second, health status may be endogenous, and non-participation in the labour force may affect health (either positively or negatively). Regarding the first criticism, Bound and Burkhauser (1999) survey a number of papers that show that self-reported measures are highly correlated with *clinical* measures of disability. Benitez-Silva et al. (2004) show that self-reports are unbiased predictors of the definition of disability used by the SSA. 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

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

between 1983 and 1996 that are not significantly different from the employment trends for the broader population of people with an impairment. See however Kreider (1999) and Kreider and Pepper (2007) for evidence based on bound identification that disability is over-reported among the unemployed. Table 1 adds to the evidence in support of our self-reported measure of work limitation by using the 1986 PSID health supplement to show how objective measures of limitation vary with the self-reported status (sampling weights are used throughouts). This correlation reinforces the evidence from Burkhauser and Daly (1996), who also use the 1986 supplement but a different definition of disability.

Table 1: Validity of Self-Reported Disability Status

Objective indicator	*No disability* $L=0$	*Moderate* $L=1$	*Severe* $L=2$
Trouble walking/climbing stairs	7%	58%	79%
Trouble bending/lifting objects	4%	45%	75%
Unable to drive car	0%	9%	33%
Trouble with eyesight	2%	5%	16%
Need travel assistance	0%	2%	27%
Need to stay inside	0%	5%	28%
Confined to chair/bed	0%	5%	26%
Limited in physical activity	12%	80%	94%
Spent some time in hospital	5%	24%	35%
Average # of days in hospital	0.36	1.78	14.49

Notes: The sample is male heads of household, aged 23-62. Data refer to 1986.

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

Disability Insurance To identify whether an individual in the PSID is receiving disability insurance, we use a question that asks whether the amount of social security payments received was due to disability.[19] This question is asked from the 1986 wave onwards. Prior to 1986, the question was not targeted to the head of the household, and so we cannot distinguish the recipient of the insurance.

[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.

Consumption Data One difficulty with the PSID is that the consumption in the data refers only to food. By contrast, in the model, the budget constraint imposes that over the lifetime, all income is spent on (non-durable) consumption. To compare consumption in the model to consumption in the data, we 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 CEX sample we construct to do the imputation of consumption tries to mimic as closely as possible the sample selections we impose in the PSID. Hence, our CEX sample includes only families headed by a male, reporting data between 1986 and 1992, with no missing data on the region of residence, aged 23 to 62, not self-employed, reporting data for all interviews (so an annual measure of consumption can be constructed), with complete income response, non-zero consumption of food, and not living in student housing.

We estimate in the CEX the following regression:

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

where F is food consumption.

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

We next define in the PSID the imputed value:

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

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

Sample Statistics Table 2 reports some sample statistics for individuals by work limitation status and by education (using sampling weights throughout). Regardless of education, the disabled are older, less likely to be married or white, with a smaller family, less likely to be working, and more likely to be on DI. Their family income, wages, and food spending are lower, but income from

transfers (both private and public transfers) is higher. The high educated have higher participation rates and lower DI recipiency rates.

Table 2: Sample Statistics by Work Limitation Status

	Low Education			*High Education*		
	$L=0$	$L=1$	$L=2$	$L=0$	$L=1$	$L=2$
Age	40.28	44.80	48.81	39.46	42.69	46.07
% Married	0.79	0.84	0.69	0.77	0.72	0.61
% White	0.84	0.90	0.80	0.91	0.92	0.76
Family size	3.01	3.16	2.61	2.92	2.70	2.57
Family income	43,912	39,715	26,416	66,945	51,728	36,098
Income from transfers	1,758	4,667	10,284	1,637	4,700	11,358
% Working now	0.90	0.71	0.15	0.94	0.77	0.44
% Annual wages > 0	0.97	0.81	0.19	0.98	0.89	0.48
Hours|Hours>0	2,140	1,941	1,358	2,228	2,039	1,742
Wages|Hours>0	29,618	24,518	14,718	45,713	33,447	28,365
Hourly wage	12.64	11.78	9.33	19.33	15.60	14.68
% DI recipient	0.01	0.08	0.52	0.00	0.03	0.31
Food spending	5,352	5,223	4,198	6,232	5,738	5,223
N	9,112	784	635	8,003	415	171

Notes: monetary values are in 1992$; the sample is male heads of household, aged 23-62.

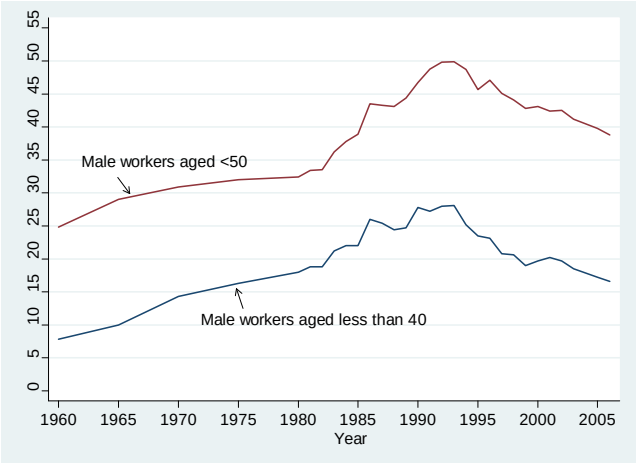


Figure 1: Proportion of new DI awards

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, the HRS samples only from a population of older workers

and retirees (aged above 50). This is an important limitation because the high current levels of DI were associated with sharp increases in the flow-on rates for the under-50s: Figure 1 shows that male workers younger than 40 account for 20 to 25% of new entrants in the Disability Insurance program in recent years, and between 40% and 50% of new entrants are under 50. We use the PSID to understand this behavior because it samples individuals from all ages and follows them across their life-cycle. The SIPP has the advantage over the PSID of being a much larger data set, but it lacks any consumption data. This is problematic because an important element of our model is the state dependence in utility induced by health.

4 Identification

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.

4.2 The Wage Process

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

$$\ln w_{it} = X'_{it}\alpha + \varphi_1 L^1_{it} + \varphi_2 L^2_{it} + \varepsilon_{it} + \omega_{it} \tag{6}$$

with $\varepsilon_{it} = \varepsilon_{it-1} + \zeta_{it}$ as before. We make the assumption that the two errors ζ_{it} and ω_{it} are independent. Based on evidence from e.g., Bound and Krueger (1995), we assume that the measurement

error ω_{it} may be serially correlated (an MA(1) process). Our goal is to identify the variance of the productivity shock σ_η^2 as well as 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:

$$\begin{aligned} P_{it}^* &= X_{it}'\gamma + \delta_1 L_{it}^1 + \delta_2 L_{it}^2 + \theta G_{it} + \vartheta_{it} \\ &= s_{it} + \vartheta_{it} \end{aligned} \tag{7}$$

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}. We assume that

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

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

$$\begin{aligned} E\left(\ln w_{it} | P_{it}^* > 0, X_{it}, L_{it}\right) &= X_{it}'\alpha + \varphi_1 L_{it}^1 + \varphi_2 L_{it}^2 + E\left(\varepsilon_{it} | P_{it}^* > 0, X_{it}, L_{it}\right) \\ &= X_{it}'\alpha + \varphi_1 L_{it}^1 + \varphi_2 L_{it}^2 + \sigma_{\varepsilon\vartheta}\lambda\left(s_{it}\right) \end{aligned}$$

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

$$\ln w_{it} = X'_{it}\alpha + \varphi_1 L^1_{it} + \varphi_2 L^2_{it} + \sigma_{\varepsilon\vartheta}\lambda\left(s_{it}\right) + v_{it} \tag{8}$$

only on the sample of workers, and with $E\left(v_{it}|P^*_{it} > 0, X_{it}, L_{it}\right) = 0$ (Heckman, 1979). The resulting estimates of φ_1 and φ_2 should be interpreted as the estimates of the effect of work limitations on *offered* wages.

4.3 Productivity Risk

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

$$g_{it} = \Delta\left(\ln w_{it} - X'_{it}\alpha - \varphi_1 L^1_{it} - \varphi_2 L^2_{it}\right) \tag{9}$$

From estimation of α, φ_1 and φ_2 described above we can construct the "adjusted" residuals (9), and use them as they were the true adjusted error terms (MaCurdy, 1982). Assuming for simplicity of notation that ω_{it} is i.i.d., we can then identify the variance of productivity shocks and the variance of measurement error using the following moment restrictions:

$$\begin{aligned}
E\left(g_{it}|P_{it}=1, P_{it-1}=1\right) &= \rho_{\zeta\vartheta}\sigma_\zeta\lambda\left(s_{it}\right) &(10)\\
E\left(g^2_{it}|P_{it}=1, P_{it-1}=1\right) &= \sigma^2_\zeta\left(1-\rho^2_{\zeta\vartheta}s_{it}\lambda\left(s_{it}\right)\right)+2\sigma^2_\omega &(11)\\
-E\left(g_{it}g_{it+1}|P_{it}=1, P_{it-1}=1\right) &= \sigma^2_\omega &(12)
\end{aligned}$$

(see Low, Meghir and Pistaferri, 2010). Here $\rho_{\zeta\vartheta}$ denotes the correlation coefficient between ζ and ϑ (which is not of direct interest). Standard errors are computed with the block bootstrap.

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 ($\pi^{Young}_{L=0}, \pi^{Young}_{L=1}, \pi^{Young}_{L=2}, \pi^{Old}_{L=0}, \pi^{Old}_{L=1}, \pi^{Old}_{L=2}$, and P^{Re}) will be achieved by Indirect Inference (see Gourieroux et al, 1993; Smith, 2006). Indirect inference is a simulation-based method that is used when the relevant likelihood function has no analytical expressions. This is indeed the case for our complex theoretical model. The difference between indirect inference and other methods based on simulations (such as Simulated Method of Moments) is that indirect inference relies on 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 process. The key idea behind indirect inference is that the parameters of the auxiliary model

are related (through a so-called binding function) to the structural parameters of interest. The latter are estimated by minimizing the distance between the parameters of the auxiliary model estimated from the observed data and the parameters of the auxiliary model estimated from the simulated data. 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.

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

The Indirect Inference statistical criterion that we use is:

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

where $\widehat{\alpha}^D$ are the moments in the data, $\widehat{\alpha}^S (\phi)$ are the corresponding simulated moments (which we average over S simulations) for given parameter values ϕ. The function $\alpha (\phi)$ is the binding function relating the structural parameters to the auxiliary parameters, and Ω is the weighting matrix. The optimal weighting matrix is the the inverse of the covariance matrix from the data, $\widehat{\Omega} = var \left(\widehat{\alpha}^D \right)^{-1}$.[20]

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

$$var \left(\widehat{\phi} \right) = \left(J' \mathbf{\Omega} J \right)^{-1} J' \mathbf{\Omega V \Omega} J \left(J' \mathbf{\Omega} J \right)^{-1}$$

where $J = \frac{\partial \widehat{\alpha}^S (\phi)}{\partial \phi}$, and $V = var \left(\widehat{\alpha}^D - \widehat{\alpha}^S \left(\widehat{\phi} \right) \right)$. Asymptotically, V reduces to $\left(1 + \frac{1}{S} \right) var \left(\widehat{\alpha}^D \right)$, but when we present standard errors, we calculate V using the simulated moments explicitly. We calculate J by finite difference.

In what follows we discuss the mapping between structural and auxiliary parameters.

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

4.4.1 Moments: Consumption Regression

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

The second possible effect of disability on consumption is through non-separabilities in the utility function. For example, if being disabled 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 nonseparability or through the income loss directly.

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

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

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 $(\alpha_3 + \alpha_4)$, and so $(\alpha_3 + \alpha_4)$ captures the preference effect induced by nonseparability.[21] The split between α_3 and α_4 is clear when insurance is full. More generally, if insurance is partial, then $(\alpha_3 + \alpha_4)$ captures both the non-separable part and the lack of full insurance for those receiving DI. However, the degree of partial insurance through DI depends on permanent income and age through the AIME formula. Indirect inference exploits this identification intuition without putting a structural

[21] 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 budget constraint effects. However, if we could find a group of individuals who are fully insured against disability shocks, then the consumption response to disability 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.

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.[22] We control for unearned income to compare individuals with the same potential for self-insurance.

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

4.4.2 Moments: Participation over the Life-Cycle

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

$$p_{it} = \sum_{x=1}^{X} \varphi_x \mathbf{1}\left\{age_{it} \in x\right\} + \upsilon_{it}$$

where p_{it} is an indicator for whether a person is working at age t, x denote the age bands and there are overall X age bands (we use four 10-year age bands: 23-32, 33-42, etc.). The moments we use as auxiliary parameters are the φ_x estimated separately for the three work limitation groups, so there are $X \times L = 12$ auxiliary parameters 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. Frictions are identified by average labor market participation and unemployment duration over the life cycle. Unemployment rates in the first periods of the 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).

4.4.3 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

[22] 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.

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 programme. Of course, in practice, the fraction who apply depends on the probability of acceptance and this is why we need to use our model to identify the actual probability of acceptance rather than just taking the observed fractions on DI as the probabilities of acceptance. For both sets of moments, we condition on being younger or older than age 45.

5 Results

5.1 Disability Risk

Figure 2 plots selected $\Pr(L_{it} = j | L_{it-1} = k)$.[23] These are transition probabilities that 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 2 plots $\Pr(L_{it} = 0 | L_{it-1} = 0)$, i.e., the probabilities of staying healthy. This probability declines over the working part of the life cycle from 0.97 to about 0.92 for the high educated and more rapidly, 0.96 to 0.88, for the low educated. The decline is equally absorbed by increasing probabilities of transiting in moderate and severe work limitations. The top right panel plots the latter, $\Pr(L_{it} = 2 | L_{it-1} = 0)$. This probability increases over the 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

[23] 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.

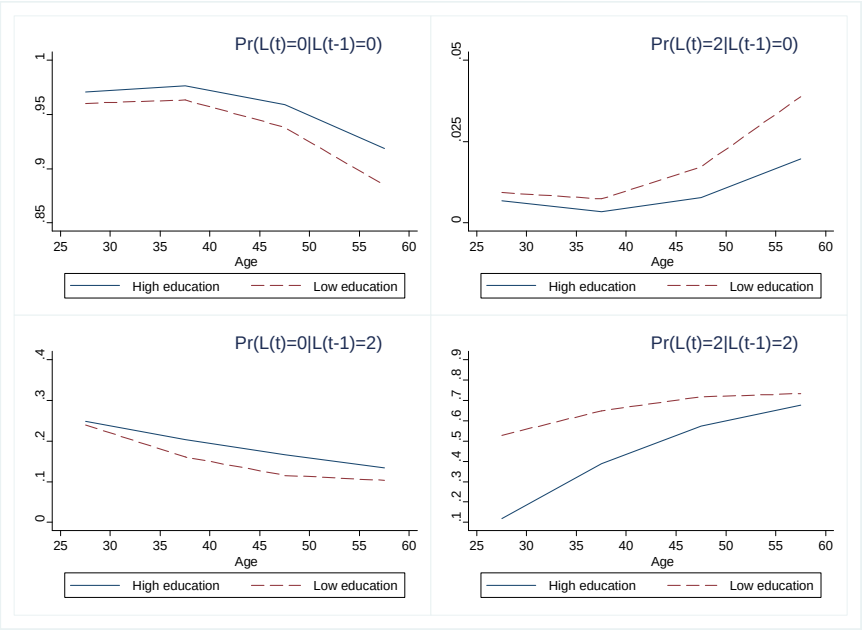


Figure 2: Selected (smoothed) Markov transition probabilities $\Pr\left(L_{it} = j | L_{it-1} = k\right)$, by education.

educated, are the reasons why we focus our remaining analysis on the subsample of individuals with low education (high school degree or less).

5.2 Wage Process

In Table 3 column (1) we report the results of estimating a simple probit regression for participation. Participation is monotonically decreasing in the degree of work limitations. We report marginal effects. Thus, the interpretation is that among the low educated, the probability of working declines by 13 percentage points at the onset of a moderate work limitation, and by 55 percentage points at the onset of a severe work limitation. As for our exclusion restrictions, their signs are correct (higher income from transfers and a more generous welfare system should increase the opportunity cost of work), and the effects are statistically significant.[24] The other effects have signs that are consistent with previous evidence.

In columns (2) and (3), 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

[24]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.

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).

Table 3: The log wage equation

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

5.3 Productivity Risk

We use the residuals of the wage equation to estimate the variance of permanent productivity shocks as well as the variance of measurement error (and the MA(1) parameter, which turns out to be statistically insignificant), allowing for endogenous selection into work (expressions (10)-(12)). The results are in Table 4. The numbers are 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 4: The variances of the productivity shocks

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

5.4 Estimates from Indirect Inference

First, we set some parameters to realistic values, as shown in Table 5:

Table 5: Exogenous Parameters

Parameter	Value	Parameter	Value
γ	1.5	T^W	160 (40 years)
R	0.016 (Annual)	T^R	40 (10 years)
β	0.025 (Annual)	λ	0.73 (Quarterly)

Ideally, we would identify the value of λ by using durations of unemployment by disability status. However, there are substantial censoring problems, as well as a large amount of noise when we stratify by education and work limitation status, and hence we take the value of λ from Low, Meghir and Pistaferri (2010).

Next, we present results on the moments we have matched by Indirect Inference. The first set of moments, in Table 6, come from matching DI policy moments: the work limitation status of DI recipients (Panel A) and the DI status of people with different work limitations (Panel B), separately for younger (age$<$45) and older workers (age$\geq$ 45). Our model is capable of matching most 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.

Table 6: Disability Insurance Moments

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

Table 7 reports the second set of moments, obtained from estimating the auxiliary log consumption equation (using imputed data, as detailed above).[25] We obtain a good match between data and simulations. The signs and in most cases even the magnitude of the coefficients are similar. These numbers are not intrinsically interesting, however. It is their link with structural parameters that it is more interesting for our purposes.

Table 7: The Log Consumption Equation

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

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

Finally, Table 8 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.

[25] 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).

Table 8: Labor Market Participation by Disability Status

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

In Table 9 we report the Indirect Inference parameter estimates corresponding to these moments, obtained by minimizing the distance between the moments computed from the data (i.e., those reported in Tables 6, 7, and 8), and the equivalent moments computed from the simulated model. We estimate that a moderate (severe) disability induces about a 4% (8%) loss of utility in terms of consumption. Participation induces a 32% loss.[26] 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 discusse further in the next section.

Table 9: 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)

[26] 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).

Note: 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.

5.5 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. However, we reproduce in table 10 the main flow statistics and the simulated counterparts as an indication of the performance of the model.

Table 10: Flows onto and off Disability Insurance

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

5.6 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 Award rate: $\Pr(DI = 1 \mid DI^{App} = 1)$. We estimate this rate (using our structural model and estimated parameters) to be 0.40. During the period covered by our data (1986-92), there were 3.3 million awards made to 7.8 million applicants, resulting in a 42% average success rate.[27] Our estimate contrasts quite well also with the reduced form estimates (0.45) obtained by Bound and Burkhauser (1999) and others using data on individual DI application and DI receipt from the HRS.

Given that the true disability status of an applicant is private information, SSA evaluators are bound to commit two types of errors: Admitting into the DI program undeserved applicants and rejecting those who are truly disabled. Our estimates show how large are the probabilities associated with these errors. Consider first the extent of false positives (the proportion of healthy

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

applicants who receive DI). From Table 8, 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: $Pr(L = \{0, 1\}|DI = 1, DI^{App} = 1) = 0.10$. In the literature, we have found reduced form estimates that are fairly similar, 0.16 to 0.22 in Benitez-Silva et al. (1999), depending on the statistical assumptions made, and 0.19 in Nagi (1969).

Consider next the probability of false negatives (i.e., the proportion of severely disabled who apply and do not receive DI). From Table 8, we estimate that the type I errors are 65% for the younger and 28% for the older workers. The fraction of 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 0.52-0.60, and Nagi (1969), 0.48. These comparisons confirm that our structural model is capable of replicating quite well reduced form estimates obtained using direct information on the application and award process. Our estimated award process is slightly more efficient than previous estimates, but the differences are slight.

Finally, with an estimated reassessment rate of 5%, we predict that an individual on DI is expected to have his disability status reviewed approximately every 20 quarters.[28] 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%.

6 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 programs. We consider four changes: first, making the program "stricter" by increasing the threshold that needs to be met in order to qualify for benefits; second, changing the generosity of disability payments; third, changing the reassessment rate of disability recipients; and finally, we consider changing the generosity of the food stamp programme. 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

[28] 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.

all ages which makes the individual indifferent between the status quo and the policy change considered.[29] The policy changes we consider generate behavioral effects, such as changes in labor supply and savings. In all the experiments below the impact on the government budget is neutralised by adjusting the wage tax iteratively using equation (5).

6.1 Strictness of DI Admissions

Increasing the strictness of DI admissions has been advanced as one possible solution to the incentive problem. Increases in strictness in 1980 led to sharp declines in inflows onto DI, although the criteria was relaxed again in 1984. The issue is whether the benefit of improved incentives outweighs the worsening insurance. To tackle this issue, we need first to define a 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:[30]

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

Figure 3 illustrates the resulting distributions of S for those over 45 by work limitation status.

[29] This is obtained by calculating expected utility at the start of the life-cycle before the resolution of any uncertainty.

[30] 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. The advantage of the Beta distribution is that the precision of the signal increases as true disability status worsens.

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

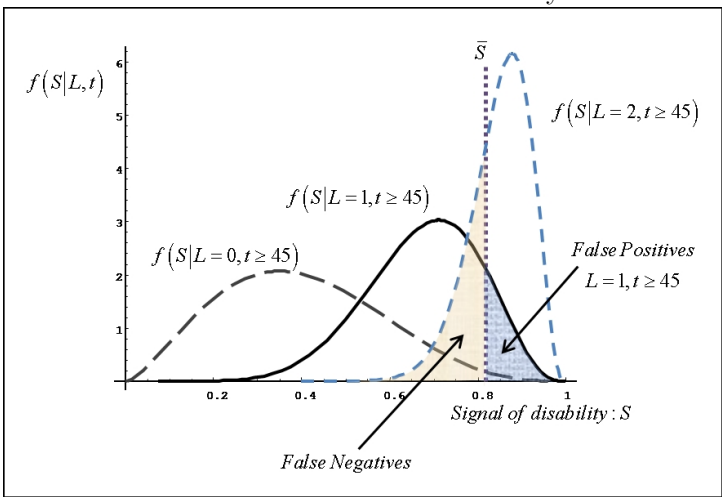


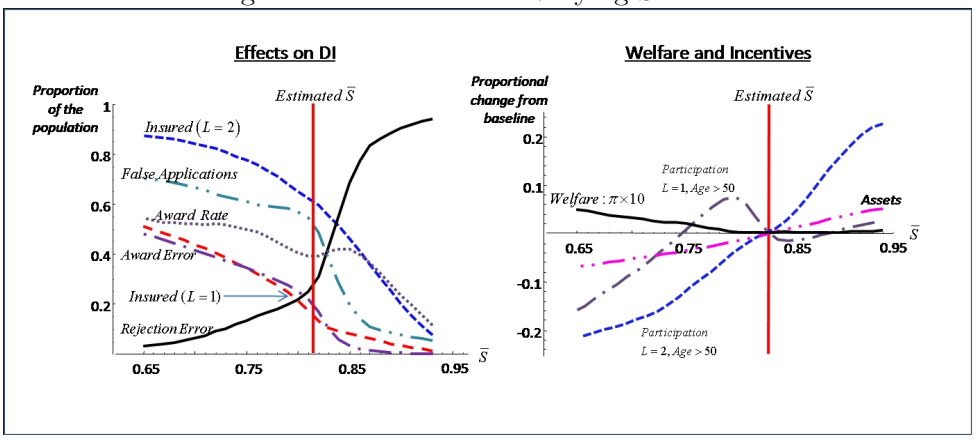
Figure 3 illustrates some of the errors under the estimated DI program. The area on the left of $\overline{S}$ under the dashed light grey curve (labeled $f(S|L=2,t\geq 45)$) measures the probability of rejecting a deserving DI applicant. The area on the right of $\overline{S}$ under the solid grey curve (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 insurance, and we use our model to determine which dominates when the strictness of the test changes.[31]

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

[31] 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.

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 $(L = 2)$").

Figure 4: The Effects of Varying Strictness



The right hand graph shows the incentive effects of the alternative $\bar{S}$, as well as the willingness to pay. For all variables considered, the y-axis measures the proportional change relative to the baseline.[32] 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, and so participation can fall as strictness increases through this indirect mechanism. The effects on participation for those who are not work limited at all are negligble.

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 0.05

[32]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.

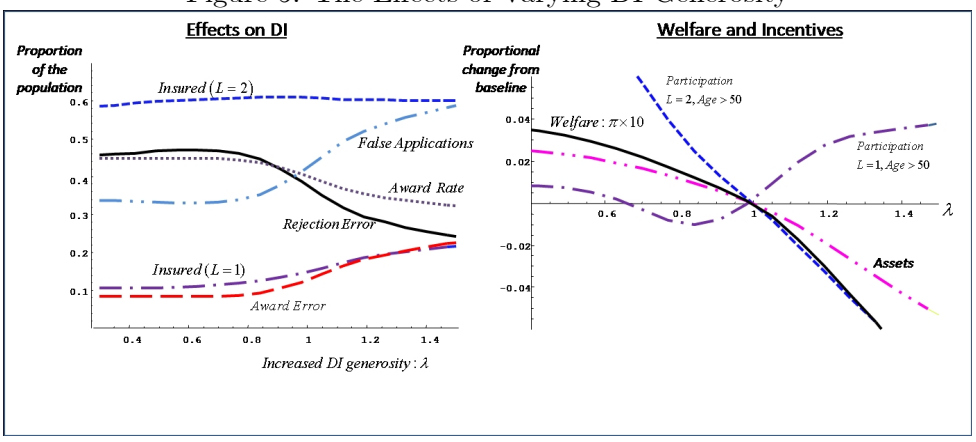
(0.5%). This gain is the net gain of two offsetting effects: there is a benefit of increased insurance against disability which individuals are willing to pay for, but this is partly offset by a loss arising from output being lower as individuals work less. Part of the benefit of the relaxed strictness 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.[33]

6.2 Generosity of DI Payments

Figure 5 shows the effects of proportional changes in DI generosity, with the proportional changes ranging from a cut to 30% of its current value to a 50% increase. The budget impact of all changes are neutralised by adjusting the wage tax iteratively using equation (5).

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.

Figure 5: The Effects of Varying DI Generosity



[33] We have considered various alternative specifications for the distribution of the noise over 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.

Given these effects, the welfare implications of changing generosity shown in the right hand graph of Figure 5 are not surprising: 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.

6.3 Reassessment of DI Recipients

In Figure 6, we consider changing the reassessment rate. Given our estimate of the cost per reassessment, 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 false applications from 54% to 30%. This in turn leads to a decline in the award error, and a decline in the fraction of the non-work limited who receive insurance. For the moderately disabled who are 45 or over, the decline is from 24% to 10%. 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.

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 0.043%. Increased reassessment increases participation among the severely work limited, who are discouraged from applying or removed from the DI rolls. This also leads to greater saving, which discourages participation, particularly among the moderately limited.

6.4 Generosity of The Food Stamp Program

Figure 7 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,

Figure 6: The Effects of Varying Reassessment Rates

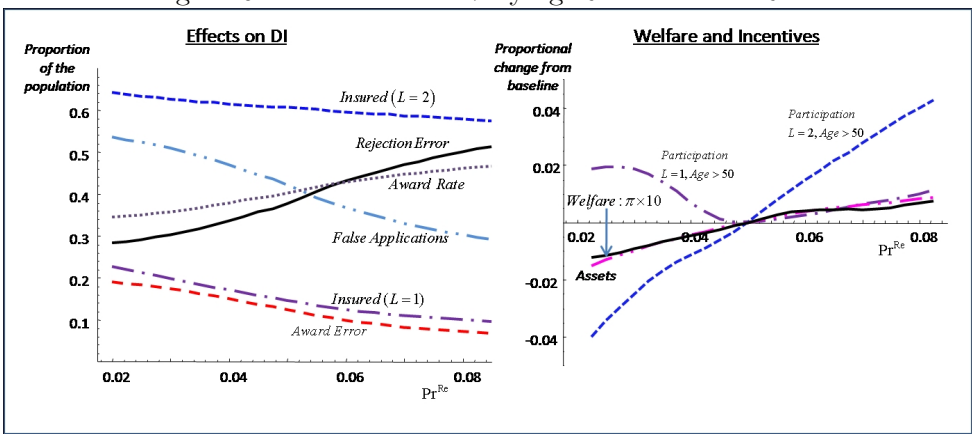
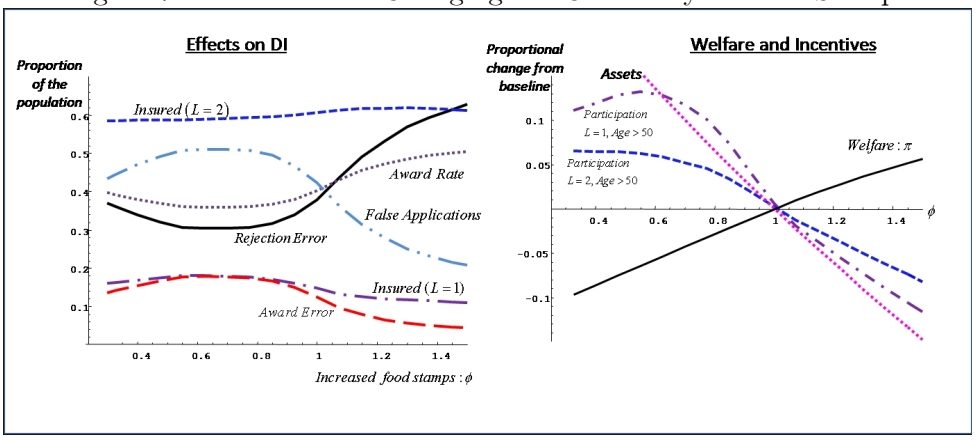


Figure 7: The Effects of Changing the Generosity of Food Stamps



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 are sufficiently generous that false applications for DI fall. This effect translates into the fraction of those not severely disabled who are in receipt of DI (the "Insured ($L = 1$)" line shows this for those 45 and over) and into the award error: both of which decline 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 productivity with no risk of rejection. Together, these effects imply substantial welfare increases as the generosity of food stamps, funded by a wage tax, 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.

7 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. These are intended to illustrate the trade-offs from the various policy options. Increasing the strictness of the screening process through increasing the work limitation threshold for qualification reduces the number of individuals receiving benefits among both the severely work limited and among the healthy because of the noisiness of the signal of work limitation status. Thus increased strictness leads to a decline in welfare because the existing program already suffers from turning down large numbers of severely disabled. For other reforms, 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. Of course, the severely disabled will then receive less insurance, but this change increases welfare ex-ante. Similarly, increasing the generosity of Food Stamps leads to a fall off in false applications for DI and mis-directed insurance, leading to better targetting of DI and a welfare improvement. More frequent reassessments of recipients directly reduces the number of claimants who are not severely work limited, but equally importantly more frequent reassessments substantially reduce the proportion of false applicants. This leads 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. The conclusions arose because welfare improving reforms lead to a separation of the severely work limited from the moderately limited for whom work is a realistic option. One difficulty with this conclusion is the clear non-linearities in behaviour apparent from the simulations in section 6.

In terms of extensions, our model of the disability insurance process is incomplete: Benitez-Silva et al. (2004) have emphasised 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.

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

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

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

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

[8] Benitez-Silva, Hugo, Moshe Buchinsky, and John Rust (2006b), "Induced Entry Effects of a $1 for $2 Offset in SSDI Benefits", unpublished manuscript.

[9] 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.

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

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

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

[13] 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

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

[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] Daly, Mary C. and Richard 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.

[18] Danforth, J. P. (1979) "On the Role of Consumption and Decreasing Absolute Risk Aversion in the Theory of Job Search" in: Studies in the Economics of Search, ed. by S. A. Lippman, and J. J. McCall, pp. 109-131. North-Holland, New York.

[19] DeLeire, Thomas (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, Oliver and Jean-Baptiste Michau (2010), "Optimal Social Security with Imperfect Tagging", mimeo.

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

[23] 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.

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

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

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

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

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

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

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

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

[32] Lentz, R. and T. Tranaes (2005), "Job Search and Savings: Wealth Effects and Duration Dependence", Journal of Labor Economics, 23(3), 467-490.

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

[34] Low, Hamish, Costas Meghir and Luigi Pistaferri (2010), "Wage risk and employment risk over the life cycle", American Economic Review, forthcoming.

[35] 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.

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

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

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

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

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

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

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

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

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

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