Disability Insurance: Theoretical Trade-Offs and Empirical Evidence

Hamish Low[*] Luigi Pistaferri[†]

April 8, 2019

Abstract

Disability insurance provides protection following health shocks that limit the ability to work. We discuss the traditional trade-off between insurance and incentives in providing this insurance, with a focus on the US and UK experiences. Disability insurance programs are large and growing, both in expenditure and in numbers of recipients. This growth may be due to declining opportunities for people at the bottom of the wage distribution, to declining health in the working-age population, and to the increasing recognition of mental and muscolosketal disorders as work-related disabilities. We provide a life-cycle framework for analysing the trade-offs and discuss recent empirical evidence on the different sides of the insurance-incentive trade-off and of the efficiency of the program. There is substantial evidence on the extent of labour supply incentive costs of disability insurance, but there has been a lack of evidence on the insurance value until very recently. Further, evidence on errors in the disability insurance process suggests false rejections of genuine claimants is a substantial problem, and these are more serious than false acceptances of healthy applicants. The underlying point is that reform should be focused on reducing false rejections and supporting labour market attachment, for example through allowing partial disability payments and rehabilitation, rather than on reducing false positives. The difficulty in considering reform is that the design of disability insurance has many aspects that all interact and impact on outcomes.

Keywords: social insurance, incentives, false rejections

[*] Oxford University and IFS.

[†] Stanford University, SIEPR, NBER and CEPR.

1 Introduction

Disability insurance is a key part of social insurance provision across the OECD. It provides insurance against extreme health shocks that prevent individuals from working. The scale of these programs has been growing fast, both measured in terms of total spending and in the fraction of the population in receipt of benefits. Compared to analysis of unemployment insurance, the economics literature has been slow to engage with understanding the scale of the programme or reaching consensus about the incentive and insurance value of the programme. Much of the current debate about the scale of the programs focuses on the incentive costs, particularly in terms of reduced labour supply. On the other hand, and less researched, there is a substantial value of these programs to those in severe need. A key conclusion from our reading of the evidence is that the focus of the literature on the incentive costs and extent of false claimants paints an unduly negative picture of disability insurance. By contrast, as we discuss in terms of future work and policy reform, the focus should be on how to improve the insurance targetting, to reduce false rejections and how to improve labour force attachment, rather than on how to reduce false applications per se.

There is little evidence that health is declining in the general population,[1] and so the growth in disability insurance programs has to be explained by changes to the eligibility rules, as well as changes in incentives and in opportunity costs. The open question is how much of this growth can be attributed to genuine declines in health, at least for some population groups; how much to the increasing recognition of mental and muscolosketal disorders as work-related disabilities; and how much to acceptances onto the program by those who are healthy or looking for a substitute for unemployment insurance or retirement. At the heart of these questions is the issue of the trade-off between providing coverage for individuals who are genuinely in need, and avoiding giving benefits to those who are healthy and able to work. The aim of this survey is to provide a framework for thinking about this trade-off and to discuss recent empirical evidence on the different sides.

In the presence of disability insurance, increased benefit generosity or higher opportunity costs of working create incentives for people to leave work and apply for the program; for staying on the program even when their health status improves; and in extreme cases, to exaggerate their disability in order to be awarded benefits. We think of individuals as having a level of productivity that depends on health and skills. Some individuals with a moderate disability may choose to apply

[1] The exceptions being the obesity crisis (Lakdawalla, Bhattacharya and Goldman, 2004) and the opiod crisis in the US, which increasingly affect the young, and have contributed to increasing mortality rates in these groups after a secular decline, and to a decline in exits from disability insurance rolls (Morden et al., 2014).

for DI applicant if their skills deteriorate due to external shocks, such as automation, international trade, etc. The issue is that the "true" disability status of an individual is unobserved and the screening naturally imperfect. The screening process in DI is in fact much more difficult than in unemployment insurance, where the only issue is whether people voluntarily quit or are laid off: a screening decision that a cooperating third party (the firm) typically helps to resolve with little to no error. In the case of DI, the screening process is instead prone to errors: rejecting a truly disabled person (Type I error), as well as making an award to someone who is not truly disabled (Type II error).[2] Most of the difficulties involved in the screening process come from the fact that a disability involves a mixture of medical, psychological, and social difficulties, and it may be extremely hard to make a correct decision even with rich medical information. This is especially true when the decision has to be or the reject/award type (as in the US) rather than deciding how fractionally disabled a person is (as in Italy or in the Veteran Disability program). Another difficulty is that given the low exit rates from the program, disability insurance acts effectively as insurance against permanent shocks: it meets demand for long-term protection against long-term unemployment or productivity declines that cannot be satisfied by other social insurance program (such as UI) which are temporary by design. It becomes extremely hard to distinguish people whose productivity has declined because of poor health from those who faced sharp decline in productivity which in turn led to poor health.

A broader normative issue, which we do not address here given the more narrow focus of the survey, is whether governments should introduce insurance programs against permanent shocks that condition explicitly on shocks like automation or international trade, rather than having job and wage losses due to these events being absorbed by programs that were not designed for them, such as DI. The conditioning on poor health in DI programs is to avoid providing payments for people who have a high preference for leisure. But our key conclusion that that the programs experience large Type I errors despite the conditioning, indicates that there is substantial underinsurance by the government of long term productivity shocks.

The survey is structured as follows. We start in Section 2 with a discussion of institutional details about disability insurance programs focusing mainly on the US and the UK; and discussing the different dimensions of policy design, such as the nature of the medical test and the degree of progressivity. Section 3 provides a framework for discussing the insurance and incentives that

[2] Benitez-Silva et al. (2004) have also emphasized rejection errors (the fraction of truly disabled people who are rejected) and award errors (the fraction of awarded people who are not severely disabled). Of course, these errors are connected to Type I and Type II errors through the Bayes' formula.

pervades all social insurance programs, and zooms on Disability Insurance programs as a case in which this trade-off is at its most visible. Section 4 discusses the empirical evidence on incentive effects and insurance implications. The focus in the empirical literature, perhaps because the issues are much better defined, has been heavily on incentives.[3] There has been comparatively less work on the insurance side, although in recent years the literature has become more diversified. The final, concluding Section 5 discusses policy implications.

2 Institutional Detail and Statistics

Many OECD countries offer support to those with health conditions that stop them being able to work, but the details of how the programs work and the consequences for the wider economy vary enormously across countries. We begin this section by contrasting the scale of disability programs over time and focus mainly on comparing the US and UK. We then discuss alternative aspects of the institutional structure which determine the incentive and insurance effects of the programs.

We use the term disability insurance (DI) to cover programs that offer income in replacement of earnings, where the loss of earnings is due to bad health. In the UK, this includes programs such as incapacity benefit and the employment support allowance. In the US, this definiton of DI includes the Social Security Administration's disability insurance programme, as well as Supplemental Security Income (SSI) which is means-tested rather than contributory, but still aims to replace lost or low earnings capability due to health. By contrast, we use the term disability benefits to cover programs that offer direct help with costs of being disabled whether an individual is working or not.

2.1 The Scale of Disability Programs

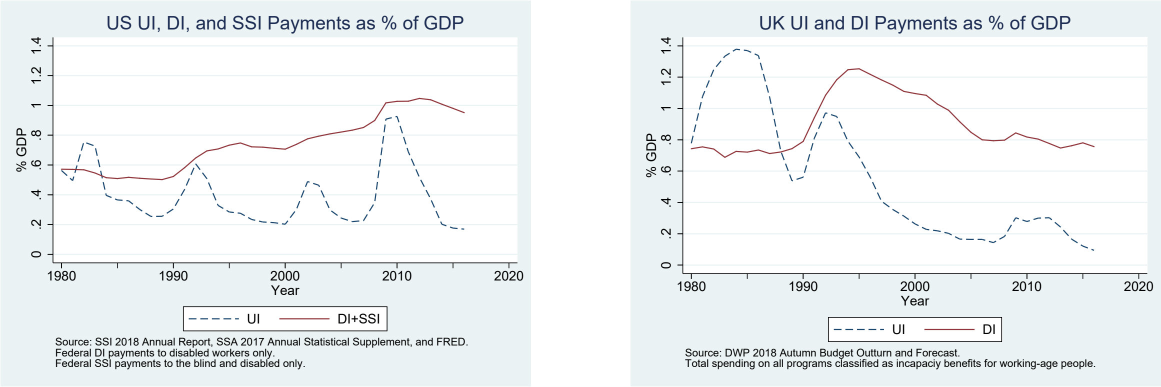
We can measure the scale of disability programs through expenditure and through the number of recipients. Figure 1 shows how spending on DI as a fraction of GDP has increased over the past 30 years in the US and UK. The figure also shows spending on UI. The first point to stress is the scale of the DI program: approximately 1% of GDP is now spent on DI in the US, and slightly more in the UK. Spending in other countries such as Norway and Sweden is even higher. With the exception of the Great Recession period, spending on DI is much larger than spending on UI.

Figure 2 shows the growth in the recipiency rates for the US and UK. In the US, 4.5% of the

[3] Cases of "disability cheaters" are also more frequently highlighted in the popular press and are likely to spark more outrage than controversial cases of denials.

working age population receive DI, and the program has been growing at a fairly constant rate since a reform in 1984 relaxed the admission criteria. In the UK in 2017, 6% of the working age population received disability insurance through the employment support allowance, although this is below the peak of over 7% in 2004 (see also Banks, Blundell, and Emmerson (2015)). The sharp rise in recipients in the UK happened at the start of the 1990s, and this led to reform in 1995. Offsetting the plateauing and then decline in claimants of disability insurance in the UK, there has been an ongoing rise in the fraction receiving disability benefits that directly support the extra financial costs of a disability. This rise is despite the reform of 2010 aimed explicitly at reducing this support.

Figure 1: Spending on Disability and Unemployment Insurance



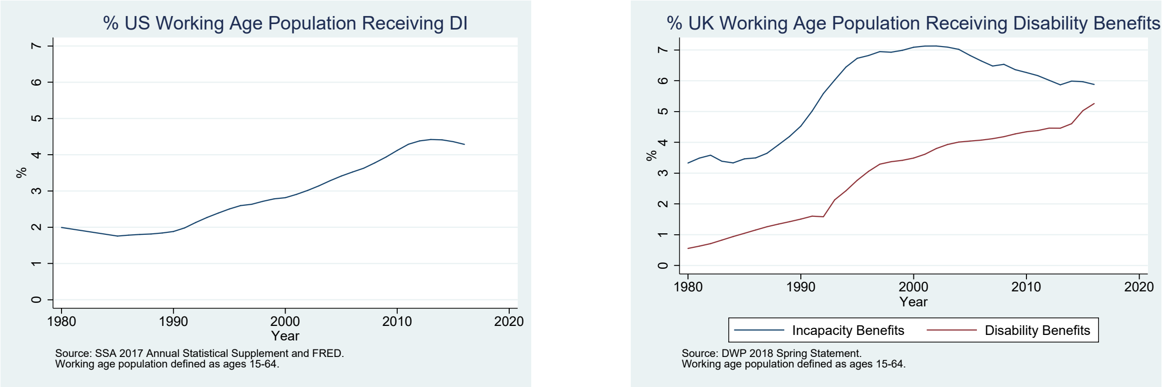
Source: SSI 2018 Annual Report, SSA 2017 Annual Statistical Supplement, and FRED.
Federal DI payments to disabled workers only.
Federal SSI payments to the blind and disabled only.

Source: DWP 2018 Autumn Budget Outturn and Forecast.
Total spending on all programs classified as incapaciy benefits for working-age people.

(US)

(UK)

Figure 2: Stock of Recipients



Source: SSA 2017 Annual Statistical Supplement and FRED.
Working age population defined as ages 15-64.

Source: DWP 2018 Spring Statement.
Working age population defined as ages 15-64.

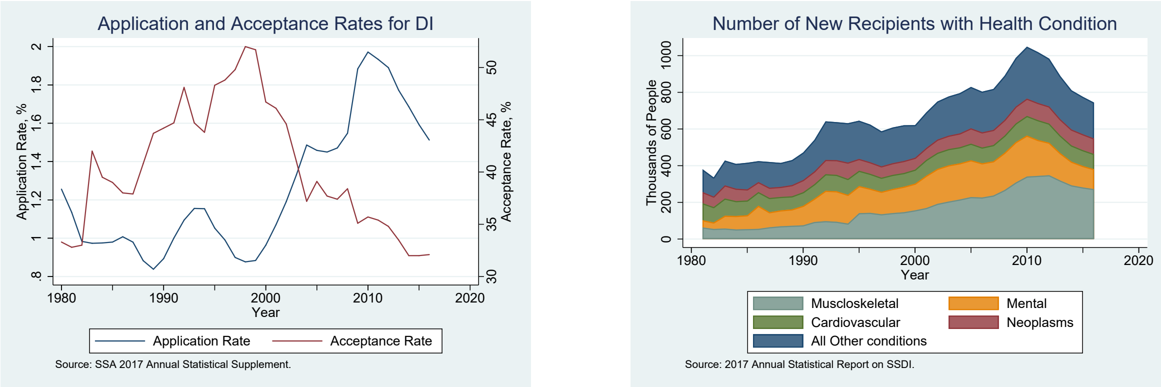
(US)

(UK)

For the US, we are able to break these numbers of recipiency down further. The left-hand graph of Figure 3 shows the evolution of the application and acceptance rates: application rates have increased sharply since 2000, and with a corresponding fall in the acceptance rate. The net effect however is an ongoing increase in the number of new recipients, as shown on the right hand side of the Figure. The figure also shows how this increase is split between different health conditions: the growth in the number of new recipients is among those with mental health conditions and muscloskeletal/backpain. This growth in new recipients translates into the large increases in the stock shown in Figure 2 because those entering with back pain or mental health problems tend to be younger and so stay on the program for longer.

Figure 4 shows what has happened to labour force participation in the US over the same time span. The striking, albeit unsurprising, point is that labour force participation rates for the disabled are much lower than for those who are not disabled. As we discuss in the modelling of disability in section 3 below, this low level of participation may reflect lower wages, higher fixed costs of work or different utility costs. It may also reflect the availability of disability insurance to protect this disabled group. Finally, it may reflect (perversely) legislation such as the American with Disability Act (ADA), see De Leire (2000) and Acemoglu and Angrist (2001). In addition to the level of labour force participation being low, there has been a sharp decline among those who are disabled. For men with a disability, participation rates have declined from 35% to a low of 18% over the past 30 years.

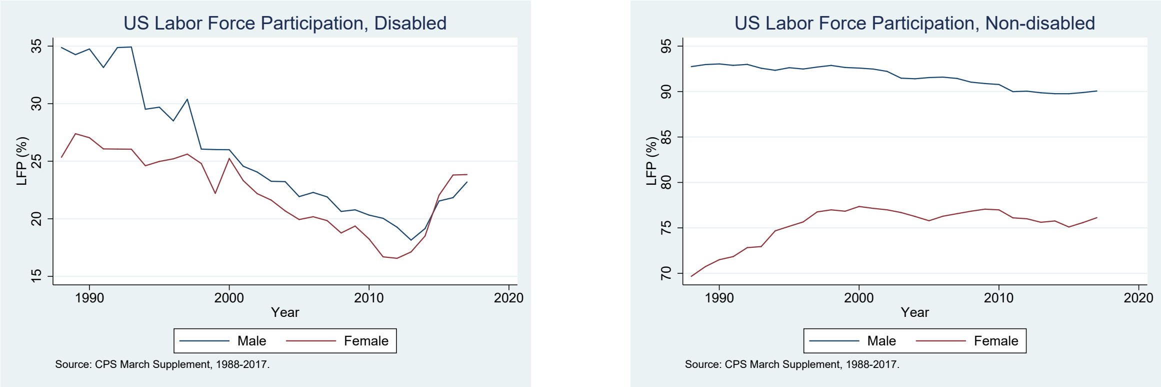
Figure 3: Application Rates and Recipiency: US



Source: SSA 2017 Annual Statistical Supplement.

Source: 2017 Annual Statistical Report on SSDI.

Figure 4: Labour Force Participation: US



2.2 Dimensions of Policy

This discussion of the evolution of DI programs and how programs differ between the US and the UK highlights the importance of understanding the diminsions of DI policy. Programs differ in five main dimensions:

1. The nature of the medical test;

2. The application process and labour market attachment during application;

3. Eligibility: hether the program is contributory, requiring individuals to have made social insurance-type payments before claiming, or means-tested;

4. The generosity and progressivity of benefits;

5. The process of reassessment and having benefits removed

We discuss these dimensions in turn. As we show in section 4, separating out the policy options into these different aspects highlights the dearth of empirical evidence on the consequences of different designs, and the need for analysis of DI programs to be capturing institutional details. Instead, the empirical evidence is on the total effect of a particular DI system.

The Medical Test

Individuals at all ages can suffer shocks to their health. The issue is when these shocks are severe enough to qualify for disability insurance. A stricter medical test may reduce false applications, but might well increase false rejections, worseing insurnace.

A narrow, strict qualifying definition is a solely medical assessment that an individual is unable to do any work at all. A broader, more lenient qualifying definition would be based on whether the health condition stopped the applicant doing certain types of work. In this case, even if other sorts of work were still feasible, the applicant would still receive DI. This more lenient definition is related to the notion of disability as being a loss of earnings capacity: a loss of earning capacity means an individual may no longer be able to do their previous job, but does not mean they are unable to do any job or unable to relocate to a lesser job.

If an individual's past occupation and training is taken into account, the absence of such an 'appropriate' or 'suitable' job might be the factor that renders the individual disabled. This generous criterion was used in the Netherlands before 1987, Italy before 1984 and Sweden before 1993. Similarly, if the criterion is a loss of earning capacity, then the disability definition is tied to a statement about local labour market conditions. The leniency of the definition of disability is further affected by varying the percentage loss of earning capacity that qualifies as being complete disability: for many individuals, a health shock may reduce their earnings capability but not prevent them from working altogther.

These broader economic definitions of disability lead to an assessment of more individuals as being disabled and to the numbers going onto disability varying with the business cycle (Benítez-Silva, Disney, and Jiménez-Martín (2010) and Black, Daniel, and Sanders (2002)). This in turn makes the distinction between unemployment insurance and disability insurance blurred.

The breadth of definition used to determine disability benefit eligibility varies across countries. The approach of the UK and the Netherlands is to concentrate on what an individual is still able to do following a health shock, using residual work capacity and earning capacity respectively. In contrast, the US defines those unable to earn more than a certain monthly threshold due to a health condition to be disabled.

There are two further crucial issues concerning medical evidence. The first concerns who actually reviews the evidence and makes the decision, and in particular whether an applicant's own doctor makes the medical case which is then processed by a non-medical administration official, or whether the social security administration appoints their own doctor. An intermediate case is where the individual's own doctor submits evidence to the administration, which is then assessed by an administration doctor. The second issue is the set of medical conditions necessary to be classified as disabled. In particular, whether such a set of medical conditions includes includes mental illness, muscular-skeletal pain, drug addiction and so on.

In the US, individuals submit written evidence from their own doctors or medical providers that is assessed by the social security administration local office. The criterion for admission is the presence of a disability defined as "*Inability to engage in any substantial gainful activity (SGA) 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.*" This is a combination of a medical criterion and an economic criterion. Despite the formal criterion changing very little, there have been large fluctuations over time in the award rates as seen in Figure 3: award rates fell from 48.8% to 33.3% between 1975 and 1980, but then rose again quickly in 1984, when eligibility criteria were liberalized, and an applicant's own physician reports were used to determine eligibility, to a peak of over 50% in 1999. Further, although in 1983, 82% of initial DI awards were made strictly on medical criteria, by 2001 this had dropped to just 58% (Autor and Duggan, 2006).

In the UK, before 1995, the medical criterion was lenient: the requirement was that the individual could not do a job that was appropriate to her qualifications and work-history. This was tightened in 1995 so that the existence of any work that the claimant could do was enough to render them non-disabled. Further, the medical assessment was no longer conducted just by personal doctors. The introduction of the Employment Support Allowance (ESA) in 2008 went further and introduced a new eligibility test, the Work Capability Assessment. This consists of two parts: the "limited capability for work" assessment which determines whether the claimant is entitled to ESA, and the "limited capability for work-related activity" assessment which evaluates whether an individual who has passed the first stage of the test is able to take part in work-related activities. These individuals are divided into the Work-related group or the Support group, where the latter are not expected to do anything to improve their chances of finding work due to a long-term incapacity.

The Application Process and Labour Force Attachment

There are two main issues in the design of the application process: first, the length of time to make an assessmenet; and second, the degree of labour market detachment during the application process.

The US program is very much targeted at individuals who are out of the labour market, and labour market detachment is required. There is a statutory five-month waiting period out of the labor force from the onset of disability before an application for DI will be processed. For those who are initially rejected, further appeals and processing mean the average time until the final acceptance decision is substantial (French and Song (2014) and Benitez-Silva, Buchinsky, Chan, Rust, and Sheidvasser (1999)). For 2006, the SSA report (cumulative) that waiting times for a decision were 131 days for the initial decision, 279 days for a reconsideration and 811 days for a decision on appeal to the Administrative Law Judges.

The underlying idea is that only the truly in need will be willing to wait and to apply. This US system is in contrast to the majority of other countries where disability insurance follows on from sickness benefit. This difference matters because in the US there is no direct transition from work onto disability insurance: those who apply for disability have to have alternative means of living or rely on means-tested benefits. By contrast, in many European countries (e.g., Sweden, Netherlands, France, Germany, the UK) the transition is from work immediately onto sickness benefits and then onto disability insurance.

At the time of a health shock, early intervention may be crucial in maintaining labour market attachment and in keeping individuals from becoming long-term benefit claimants. Despite this, several countries have adopted hands-off approaches, such as the US which offers no vocational rehabilitation when health shocks strike, in addition to requiring detachment from the labour force in order to apply. This is in contrast to Denmark, where (apart from in extreme circumstances) an individual must undergo vocational rehabilitation before becoming eligible for benefits, and Sweden, where individuals can combine vocational rehabilitation with partial disability benefit. The recent move in the UK to identify a work-capability group is to maintain engagement with working.

Eligibility: Contribution Requirements and Means-Testing

The distinction often drawn between social insurance schemes and welfare schemes is that eligibility for the former depends on prior contributions, and eligibility for the latter is on a means-tested basis. In the case of disability insurance, the same medical criteria is usually used and so the distinction determines primarily the amount of payments that are made.

In the US, DI is the contributory program that provides cash and health care benefits for covered workers, their spouses, and dependents.[4] By contrast, SSI is the means-test welfare program: applicants must have limited income and limited resources. The definition of low income and low resources is similar to the one used for other federal welfare programs, such as the Food Stamps

[4] There are two work requirement tests that individuals must pass: the "recent work test" and the "duration of work test". The "recent work test" requires that individuals aged 31+ have worked at least 5 of the last 10 years. The "duration of work test" requires people to have worked a certain fraction of their lifetime. For people aged 40+, representing the bulk of DI applications, the fraction of their lifetime that they need to have worked is about 25%.

program.[5]

Eligibility for disability insurance in the UK was based on previous work history, and specifically a history of national insurance contributions, alongside disability related supplements for those on the means-tested income support. This split between contributory and non-contributory payments was changed with the introduction of the Employment Support Allowance which explicitly had a contributory and non-contributory (income based) component.

The strictness of the means-test also varies substantially: in the UK, the asset test for income-based ESA is a maximum savings of £16,000, which is substantially higher than the $ 2000 of savings for individuals applying for SSI in the US.

Generosity and Progressivity

Generosity varies substantially across countries. Some countries offer earnings related benefits (e.g. Norway, Sweden and the US), whereas others (e.g. the UK) offer a flat-rate benefit that varies only with the duration of benefits, and these may be indexed to price inflation or average wage inflation.

For DI in the US, cash benefits are computed using the same formulae used to compute Social Security retirement benefits.[6] While benefits are independent of the extent of the work limitation, caps on the payroll tax financing the DI program as well as the nature of the formula determining benefits make the system progressive. SSI benefits are adjusted annually, but in 2017, an individual (couple) would receive $735 ($1,103) in cash benefits per month. In addition to the benefit payments, individuals who receive DI also receive Medicare benefits, but only after being on DI for two years and regardless of medical spending needs. This provision of health care, albeit delayed, means the replacement rates are substantially higher for workers with low earnings and those without employer-provided health insurance.[7]

The UK in 1995 reduced the generosity of benefits, by removing the component of benefits relating to past earnings. This change had a larger impact on those with long earnings histories. Bell and Smith (2004) exploited this difference to estimate labour force participation responses, and concluded that the reform to generosity had a significant impact reducing inflows (see also

[5] In particular, individuals must have income below a "countable income limit", which typically is slightly below the official poverty line Burkhauser and Daly (2011). SSI eligibility also has an asset limit ($2,000 for individuals and $3,000 for couples.).

[6] DI beneficiaries receive indexed monthly payments corresponding to their Primary Insurance Amount (PIA), which is based on taxable earnings averaged over the number of years worked (known as Average Indexed Monthly Earnings, or AIME).

[7] The formula for computing wages is indexed to average wage growth in the economy, which implies that an increase in wage inequality at the bottom increases replacement rates for low-wage individuals even further (Autor and Duggan, 2006).

Anyadike-Danes and McVicar, 2008).

By contrast to the US, the UK disability insurance program is now flat rate, although that flat rate differs according to the disability group. The work-related group receives the same amount as someone in receipt of job-seekers allowance, where as the support group receive about 50% more.

In addition to these payments to replace lost earnings, some programs as in Norway and Sweden offer allowances for extra costs such as the cost of care or medical expenses. The UK too provides extra income according to costs directly related to disability. No such provision is available in the US.

Reassessment Process

There are two main determinants of individuals leaving DI, apart from exiting into retirement or death. First, through reassessments; and second, through incentives to return to work.

In some countries there are well-defined reassessment periods, as in Sweden where most beneficiaries are reassessed at least every three years, while in others reassessments depend on the qualifying condition, as in the US and the UK. In practice, reassessment in many countries is very infrequent, conducted in unreliable ways (i.e. by mail) due to cost reasons, and benefit awards are essentially permanent.

In the US, DI beneficiaries have their disability reassessed periodically through Continuing Disability Reviews (CDR). By law, the SSA is expected to perform CDRs every 7 years for individuals where medical improvement is not expected, every 3 years for individuals where medical improvement is possible, and every 6 to 18 months for individuals where medical improvement is expected. In this way, the probability of reassessment depends on perceived work limitation status.

Mass reassessments of large proportions of recipients have taken place, such as in the US in 1982-1983, in the Netherlands in 1994-1995, and in the UK 2009-2012 XXX CHECK.. In the US, between 1981 and 1983 there were 950,000 'Continuing Disability Reviews' (CDRs) and out of these 40% were terminated. It was this mass reassessment rather than a change in the inflow rate which led to the fall in the stock of recipients. The subsequent policy reversal in 1984 was based on the belief that many of those removed were in fact not employable, partly because of depreciation of human capital and partly because they were suffering from debilitating conditions that were no longer admissible, such as mental illness.[8] A subsequent attempt in 1996 to reduce the number of

[8] The 1984 reform broadened the definition of disability again to include general pain, to consider multiple minor impairments as being equivalent to one more serious condition in determining disablement, and to allow mental disorders. More weight was attached to functional capabilities rather than the existence of a specific medical condition,

disability benefit claimants by narrowing the criterion to exclude alcoholism and drug addiction led to many of those excluded returning onto benefits under a different disability (Lewin Group, 1998).

Similarly to the US, the reassessment process in the UK differs depending on the scale of assessment at initial award. For those deemed to have permanent limitations and so allocated to the "support group", there is no reassessment. By contrast, individuals allocated in the Work-related activity group are expected to improe and are asked to go to work-focused interviews to improve their chances of working in the future. Further, this work-related group will be allocated a re-referral date where their entitlement to benefit will be reconsidered.

In-built incentives to return to work are more prevalent in some schemes than in others. Differences include whether a recipient is able to keep receiving benefits while returning to work, and how quickly a recipient can return onto DI if the return to work fails. In Denmark, for example, benefits are gradually phased out with earnings from work, in Italy it is impossible to have work income alongside a full disability pension, and in the US for those in receipt of disability insurance, after 9 months of earning more than the fairly low 'substantial gainful activity' threshold, benefits are withdrawn.

In 1999 in the US, a number of work incentive programs for DI beneficiaries were introduced, such as the Ticket to Work program, which allowed recipients to keep their medicare eligbility for a number of years despite returning to work.

3 Organising Framework: Insurance and Incentives

Health shocks that stop an individual being able to work are rare, but can be extremely serious: insurance should be very valuable. On the other hand, government involvement in insuring individuals against health shocks arises because of market failures, specifically because of perceived adverse selection in who would take out disability insurance. While governments can force all workers to take out disability insurance, this does not solve the insurance-incentive trade-off.

To understand the interplay between the insurance and incentive consequences, we need a framework. The difficulty with modelling disability insurance is that the details of the program vary substantially, as seen in section 2. Further, by contrast with unemployment insurance, consideration of health and health shocks can impact on individuals in multiple ways. In thinking about unemployment insurance, researchers now often use the Bailey (1978) framework, discussed by

and all of these together made the line between disability and non-disability even more ill-defined.

Spinnewyn in this issue. Meyer and Mok (2018) apply this same framework to disability insurance, but the framework is much less well suited to thinking about health shocks and the complexity of disability insurance. Instead, we believe that taking an explicit stand on the individual choices, the economic environment and social insurance is a more fruitful way to explore the trade-offs, as we pursued in Low and Pistaferri (2015).

Introducing an explicit framework involves making modelling choices about preferences, about the individual resoures and about markets and government support. Health shocks, and disability insurance against those shocks, trace through into impacts on each of these modelling components. In this section, we describe the various modelling decisions that can be made. Additionally, we discuss the substantial measurement issues to consider if one wants to bring such a model to the data. In section 4 we discuss empirical evidence on different components of the model and on the implications of the model, and in section 5 below, we use this framework for understanding policy reform.

3.1 Modelling Disability Insurance

At the heart of our framework of disability is a model of individuals decision making, of uncertainty that individuals face over their lifetimes, and of social insurance provided by the government.

Individuals make decisions over their lifetimes, about saving and labour supply, as well as decisions about applying for DI. These decisions are made in the face of the various shocks to wages, to employment and to health. The shocks may be partly insured by individuals own decisions and also by the social insurance system of the government. Options for social insurance against health shocks are limited by the imperfect verifiability of health status, which means some genuine applicants are turned down, whereas other false applicants are rejected. And the prospect of these errors affects decisions to apply and decisions on labour supply and saving: capturing the interaction between these different choices is therefore crucial to capturing the interplay between incentives and insurance.

Two underlying questions loom large. First, in what ways does health and health shocks affect our modelling of individuals preferences, wages and job opportunities; and second, in what ways does disability insurance affect our modelling of individuals' opportunity sets.

Modelling Preferences

We leave to one side the issue of how decisions within a family are made, and consider an individual maximizing lifetime expected utility defined over consumption, leisure, and work disability status. The question we ask here is how does an individual's health affect their utility.

$$U_t = E_t \sum_{s=t}^{t} \beta^{s-t} u\left(c_t, l_t, d_t\right)$$

where β is the discount factor, E_t the expectations operator over future risks conditional on information available in period t, and c, l and d are consumption, leisure, and work disability status, respectively.

The first issue is the possibility that preferences for consumption are non-separable with respect to work disability status (state-dependent utility). This is important in that a decline in consumption following a disability shock may underestimate the decline in welfare. For example, it may be optimal to have higher consumption spending when disabled even in a full insurance world (i.e., in a world in which the marginal utility of wealth is equalized in the two states of good and poor health) if consumption and poor health are complements in utility. There is no consensus in the literature about whether consumption and poor health are substitutes or complements. Lillard and Weiss (1997) find evidence for complementarity using HRS savings and health status data; Low and Pistaferri (2015) confirm this finding using PSID consumption data and self-reported disability data. On the other hand, Finkelstein et al. (2013) use health data and subjective well-being data to proxy for utility and find evidence for substitutability. Empirically, it may be hard to determine because certain goods are substitutes with poor health (such as vacations), while others are complements (such as alternative transportation services, domestic services, etc.). Further, it is not clear that some forms of expenditure give utility directly, and instead serve only to change health through some production process.

Modelling Constraints and Resources

We discuss in this subsection modelling the impact of health on private resources; and in the next, modelling the mitigating impact of government insurance. The main impact of health on private resources is through the impact on wages and earnings. Health can affect productivity and hence wages directly, and health can also affect the fixed costs of going to work, as well as job offer arrival and destruction rates.

We think of wages as determined by skills (unrelated to health) ζ_{it} and work limitations d_{it}:

$$\ln w_{it} = x'_{it}\beta + \alpha d_{it} + \zeta_{it}$$

where we have also allowed for the effect of demographics and other observable components of productivity, x_{it}. The presence of multiple sources of productivity means that some people with a mild or moderate disability may become applicants, for example if their skills have deteriorated permanently, making the opportunity cost of applying for disability low; and further, that some severely disabled people may continue to work because their non-health skills are still capable of generating high wages.

Similar effects can be produced by the introduction of labor market frictions: people with marginal disabilities who face higher job destruction rates or low arrival rate of employment offers are more likely to become applicants to the program. Further, in considering participation decisions, we need to allow for fixed costs of working that can directly be related to work limitations.

In the literature unobserved skills ζ_{it} are typically modeled as stochastic processes with a high degree of persistence (for example, as random walks). But little is known about the stochastic evolution of disabilities or work limitations. Some researchers model this as an exogenously evolving process, with the degree of persistence estimated directly from the data using transitions across different disability states (mild, moderate, severe, etc.). Other researchers assume that work limitations are determined endogenously, i.e., individuals accumulate health capital (as in Grossman (1972)) that is occasionally subject to extreme negative shocks (disability). Large health capital accumulation reduces the likelihood of a disability. In principle, there are interactions with other forms of capital. High human capital individuals may understand the benefits of exercising, non-smoking, low-fat diets, etc. more than low human capital individuals. A possible confounders is that high-education is also correlated with working in occupation and industries where the likelihood of developing a work-related disability is lower.

In these more complex models when disability evolves endogenously, consumption choices, such as of out-of-pocket health care spending, could be seen as an input in the production function of health investments which augment (or replace depreciated) health capital.

Modelling Government Insurance

One way to insure against disability risk is self-insurance through precautionary savings. However, since disability is a low-probability event with (usually) catastrophic consequences, self-insurance against disability risk is rarely if ever efficient. It is also possible that some implicit insurance is provided through added worker effects or help from relatives, etc. But a quantitatively more relevant form of insurance is reliance on government transfers.[9] The challenge is how to structure the complex links between the various social insurance programs people have access to in the US, UK and other countries. There are two main challenges in thinking about models of disability insurance: first, the programs have complex rules as documented in section 2; second, there are important interactions between different social insurance programs.

If application to the disability insurance program is modeled as an explicit choice, the elements described in section 2 must be taken into account: the waiting period in the application process, the generosity of the benefit formula, the stringency of the medical screening process, and the frequency and intensity of the re-assessment process. All of these change the decision to apply for benefits or to stay on the program. One option is to model directly the "supply side", i.e., take a stand on how SSA models the screening process (i.e., the extent of signal/noise it observes, the disability threshold it sets, the resources available for re-assessment, etc.). Another option is to take a more reduced form approach in which the applicants's success is a stochastic event. In particular, one can model the probability of an award as a function of the "type", defined by skills, age, and the extent of actual work limitations. Skills and age enter directly in the screening process (especially at the "vocational" stage), while the extent of work limitations capture both the second and third steps of the screening process (people with a listed impairment have a high probability of an award), as well as implying lower noise in the screening process. The probabilities of success by type become structural parameters to estimate.

Low and Pistaferri (2015) allow for the possibility that working-age individuals may potentially have access to two sources of social insurance and two sources of welfare. Social insurance is through unemployment insurance, which is a temporary program linked to past work with limited to no screening issues, and disability insurance, which is potentially an absorbing state, but is subject to uncertainty due to imperfections in the screening process imperfections. The welfare programs included are a basic means tested income-support program, and a means tested program such as SSI that combines the presence of disability with low income. The division of these various programs into "social insurance" programs and "welfare" programs is because of differences in contribution requirements before claiming, but in practice, all four programs provide insurance

[9] About 50% of private sector workers in the US are covered (incrementally relative to DI) against disability risk through private disability insurance (see Autor, Duggan and Gruber, 2013).

against different sort of shocks: welfare programs can be thought of as insurance against having very low productivity. It is not clear why government should put more emphasis on a lack of income due to bad health rather than a lack of income due to a lack of appropriate skills. This distinction becomes further blurred when the actual criteria for DI is examined since it too contains a requirement about not having appropriate skills for work, as discussed in section 2.

The consequences of these programs are intertwinned. For example, a more generous UI could disincentivize applications to the DI program during economic busts, as may have happened during the early phases of the Great Recession in the US. A high value of income support can be complementary to DI application for disabled people who fear the uncertainty associated with the screening process in a frictional labor market. Modeling the links between these programs correctly is important for judging the success of reforms to disability insurance program, as such reforms typically impact take-up of alternative programs.

Summary

The complexity of the disability insurance process can lead research in one of two ways: either to try to characterise carefully the opportunities and preferences associated with disability; or to try to identify particular bits of the disability insurance puzzle, such as the effects on labour supply or on application rates. Which approach is valid depends on the aim of the research. However, to reach conclusions on the trade-off between incentives and insurance and the policy implications of this trade-off requires the attention to detail of the program and the choices of individuals. In section 4 we report evidence on the two approaches.

3.2 Measurement Issues

Before considering the evidence, a key issue in modelling how health shocks and disability insurance impact behaviour is how to measure the extent of work disabilities. This is of course a problem for the government in assessing claims for disability, but it is also an issue for researchers who in assessing the effectiveness of disability insurance programs need measures of disability that are as close as possible to the one adopted by SSA, as quoted in section 2 above. In principle, disability, d is a continuous measure of disability (and in some cases, it may even be a vector, if one wishes to separate the social, psychological, and medical aspects of a disability). In practice, researchers are confronted with the problem that continuous measures of disability are unavailable, and one has instead to rely on broad subjective or self-reported disability indicators. In many US data sets (such as PSID, CPS, SIPP, HRS), respondents are typically asked a simple binary question: *"Do you have any physical or nervous condition that limits the type of work or the amount of work you can do?".* Several papers use this simple binary indicators to classify people as work disabled. Some dataset (like the PSID) asks follow-up questions to those answering "Yes", such as *"Does this condition keep you from doing some types of work?"* (possible answers are: "Yes", "No", or "Can do nothing"), and (to those who answer "Yes" or "No" to the latter), *"For work you can do, how much does it limit the amount of work you can do?"* (with possible answers being: "A lot", "Somewhat", "Just a little", or "Not at all").[10] Use of multiple questions has the advantage of allowing the construction of a definition of disability that can at least distinguish between mild, moderate, and severe disability. The distinction between severe and moderate disability enable researchers to target measures of work limitation more closely to that intended by the SSA. This should reduce the measurement error associated with using just the "Yes/No" responses associated to the simple binary question.[11]

The validity of work limitation self-reports is somewhat controversial for at least four reasons. First, subjective reports may be poorly correlated with objective measures of disability. However, Bound and Burkhauser (1999) survey a number of papers that show that self-reported measures are highly correlated with *clinical* measures of disability. Low and Pistaferri (2015) and Low and Pistaferri (2019) provide additional evidence using PSID and HRS data, respectively.

Second, individuals may over-estimate their work limitation in order to justify their disability payments or their non-participation in the labour force. Benitez-Silva et al. (2004) show that self-reports are unbiased predictors of the definition of disability used by the SSA ("norms"). In other words, there is little evidence that, for the sample of DI applicants, average disability rates as measured from the self-reports are significantly higher than disability rates as measured from the SSA final decision rules. However, Kreider (1999) provides evidence based on bound identification that disability is over-reported among the unemployed.

Third, health status may be endogenous, and non-participation in the labour force may affect health (either positively or negatively). Stern (1989) and Bound (1991) both find positive effects of non-participation on health, but the effects are economically small. Further, Smith (2004) finds

[10] In the HRS, people are asked "Do you have any impairment or health problem that limits the kind or amount of paid work you could do?", and (to those who answer "Yes"), "Does this limitation keep you from working altogether?" and "Is this a temporary condition that will last for less than three months?". Low and Pistaferri (2015) define as disabled someone who answer "Yes" to the first and second question, and "Not temporary" to the third question.

[11] An alternative way to reduce such error is to classify as disabled only those who answer "Yes" to question (1) for two consecutive years, as in Burkhauser and Daly (1996).

that income does not affect health once one controls for education (as we do implicitly by focusing on a group of homogenous individuals with similar schooling levels). Similarly, Adda et al. (2009) find that innovations to income have negligible effects on health.

Finally, self-reports of disability are subject to the issue of potential lack of inter-personal comparability. Some researchers have pioneered the use of disability vignettes (see Kapteyn et al., 2007) to tackle this problem. In this literature, survey respondents are first asked if they have a health limiting condition, and to rank it in terms of severity. Respondents are then shown several vignettes, describing the situation of hypothetical people with impairments of various severity, and asked to rank the disability of the vignettes with the same question wording with which they are asked to rank their own disabilities. Under two key assumptions - vignette equivalence (the situation described in the vignette is perceived by all respondents in the same way up to a random error), and response consistency (respondents evaluate the health of the vignette characters in the same way that they evaluate their own health) - it is possible to identify inter-personal differences in subjective disability thresholds.[12]

4 Empirical evidence on the incentive-insurance tradeoff

4.1 Incentive Effects of DI

There is an extensive reduced form literature that focuses on the incentive effects of disability insurance, and measures the effect of changes in the generosity of the DI program on labor force participation and disability insurance participation (in the form of claims or applications). There is also a small literature using structural estimation to identify parameters that can be used to assess the incentive-insurance trade-off. Below, we summarize key papers as well as more recent contributions. The interested reader is referred to Bound and Burkhauser (1999) for an extensive coverage of earlier evidence. With some notable exceptions, most of the evidence we present comes from studies of the US system, and there is a clear need to broaden our understanding. Further, evidence is for the most part on the total effect of a particular disability insurance scheme, rather than on understanding the different components of design.

[12] Future work could explore the possibility of using multiple indicators of poor health available in survey data (both objective and subjective) as loading factors in a latent health framework, similarly to what done in the education literature (Cunha, Heckman, etc.).

4.1.1 Labour Supply

In terms of estimating labor supply effects, the incentive for individuals to apply for DI rather than to work has been carried out by asking how many DI recipients would be in the labor force in the absence of the program. The key difficulty is identifying an appropriate control group (see Parsons (1980) and Bound (1989)) to contrast with the treatment group of actual DI beneficiaries.

An early attempt to tackle this issue is Bound (1989), who uses data from the 1972 Survey of Disabled and Non-Disabled Adults (SDNA) and the 1978 Survey of Disability and Work (SDW). Bound compares labor market outcomes for three groups of 45-64 years old individuals observed at least 18 months after the initial application: (a) DI beneficiaries; (b) rejected DI applicants; and (c) non-applicants. Bound (1989) finds that DI beneficiaries have very low employment rates (3%), as opposed to rejected applicants (around 30%) and non-applicants (75%). von Wachter, Song, and Manchester (2011) replicate the approach of Bound (1989) with more recent administrative data. They compare labor market outcomes for four groups of individuals two years after their initial application, separating out individuals by the level at which award was made: (a) beneficiaries allowed at the first (DDS) stage; (b) beneficiaries allowed at the appeal stage (ALJ-level); (c) rejected DI applicants; and (d) non-applicants.[13] They find that two years after application, workers aged 45-64 at the time of their 1997 application who are allowed at the first stage level have an employment rate of 18%, as opposed to 25% among those allowed at a later stage, 53% among rejected applicants, and 82% among non-applicants. The difference between those rejected and accepted is around 30%, as in Bound (1989). von Wachter et al. (2011) stress that there is heterogeneity in the response to DI, and that younger, less severely disabled workers are more responsive to economic incentives than the older groups usually analyzed. This growth in younger claimants has indeed been a key change in the composition of claimants since 1984.[14]

The heart of this strategy is a comparison between those accepted onto the program and those rejected. The main difficulty with interpreting the numbers is that rejected applicants may be rejected because they are less sick, and hence more able to work. This means any employment difference between rejected and allowed DI beneficiaries overstates the incentive cost of DI and

[13] They focus on men aged 45-64 at the time of application (to replicate the sample restriction of Bound (1989)) and on a younger sample as well (aged 30-44 at the time of application).

[14] These incentive effects have implications for aggregate unemployment. Autor and Duggan (2003) find that the DI program lowered measured US unemployment by 0.5 percentage points between 1984 and 2001 as individuals moved onto DI. This movement was firstly because the rise in wage inequality in the US, coupled with the progressivity of the formula used to compute DI benefits, implicitly increased replacement rates for people at the bottom of the wage distribution (increasing demand for DI benefits). Secondly, in 1984 the program was reformed and made more liberal (increasing the supply of DI benefits).

hence rejected applicants' employment rates are an upper bound of how many DI beneficiaries would be working in the absence of the program. In principle, one could address this issue by randomizing awards and denials into the DI program, an experiment that is of course infeasible.

Chen and van der Klaauw (2008), French and Song (2014), Maestas et al. (2013) and Wu and Hyde (2018) are various attempts to find a more credible comparison group than rejected applicants. Chen and van der Klaauw (2008) use the so-called "disability matrix": individuals who cross certain combinations of age, experience and education are more likely to qualify for DI at the vocational stage. Wu and Hyde (2018), using the same HRS-SSA linked data used by Low and Pistaferri (2019), extend the analysis ofBound (1989)and compare the post-application labor outcomes of awarded applicants and applicants who were rejected at different stages of the evaluation process. Finally, French and Song (2014) and Maestas et al. (2013) use a "judge" identification strategy. In particular, they use as a control group, workers who were not awarded benefits because their application was examined by "tougher" disability examiners (as opposed to observationally similar workers whose application was examined by more "lenient" adjudicators). The key assumption is that the allocation of cases to judges is as good as random. French and Song (2014) find that labor force participation among those who had their appeals for disability insurance allowed by administrative law judges had a 25.6 percentage point lower labor force participation rate and earned $4059 less three years after receiving benefits. Maestas et al. (2013) estimate that receiving SSDI assistance results in a 28 percentage point decrease in labor force participation and reduced earnings (between $3800 and $4600) two years after award. This average masks considerable heterogeneity in the effect, depending on the type and severity of the disability.

The problem with these estimates is that they are of the "local treatment effect" type since they only use variation at the boundary of acceptance/rejection. This is problematic for two main reasons: first, those who have applied and gone through to the appeal stage are a selected group of people; and second, the very fact of applying and the time taken may itself cause skills depreciation. Autor et al. (2018) find substantial evidence of this decay of human capital during the time it takes for an application to be processed, from initial consideration to final appeal. check amount

Autor and Duggan (2003) investigate the effects of changes in the demand and supply of DI benefits that took place in the late 1970s-early 1980s. Supply effects arise from swings in the stringency of DI; demand effects originate from the implicit increase in replacement rates induced by a combination of increasing wage inequality and the fact that benefits are indexed to the average wage in the economy. Autor and Duggan show that the reduced stringency of the DI program, starting in 1984, increased disability rolls and reduced the unemployment rate among low-skill workers despite general improvements in health. These effects of increased DI generosity on labor supply stem from workers involuntarily separating from their jobs and deciding to exit the labor force, rather than from people exiting voluntarily to claim benefits.

Bound et al. (2010) specify a dynamic programming model that looks at the interaction of health shocks, disposable income, and the labor market behavior of men. The innovative part of their framework is that they model health as a continuous latent variable for which discrete disability is an indicator. They model behavior among the old (aged 50 and over from the HRS).

There is much less evidence on the disincentive effects in the UK. However, the 1995 reduction in generosity did provide some evidence because the reduction in generosity had a larger impact on those with long earnings histories. Bell and Smith (2004) exploited this difference to estimate labour force participation responses, and concluded that the reform to generosity had a significant impact reducing inflows (see also Anyadike-Danes and McVicar, 2008).

Elsewhere, Marie and Vall Castello (2012) exploit a discontinuity that is present in the Spanish DI program: at age 55, Spanish DI recipients who are deemed unlikely to find work are eligible for a 36% increase in their DI benefits. They find that LFP decreases substantially for those who face the benefit increase compared to those who do not (with the effect ranging from 3.1 to 8.4 percentage points depending on the bandwidth used). Since the Spanish DI system does not require recipients to stop working, they argue that these effects can be interpreted as a pure income effect from DI.

4.1.2 Applications

Labour supply disincentives are clearly important, but further distortions arise through the application process itself. Applications are affected by changes in benefits, stringency of the screening process, etc, and these applications are a mix of genuine and false applications. We first report evidence on the efficiency and errors of the screening process. We then report how these applications are affected by generosity.

Errors in the screening process arise because of rejections of individuals who are actually disabled (type 1 errors), and acceptances onto the program individuals who are not disabled (type 2 errors). There is often a focus on the extent of type 2 errors, but as we now discuss, the evidence is that type 1 errors are the larger issue.

An early direct attempt to measure such errors in the US is Nagi (1969) who used a sample of 2,454 individuals who had had an initial disability determination. These individuals were then examined by an independent team of medical providers, psychologists and social workers: at the time of the award, about 19% of those initially awarded benefits were undeserving (type II errors), and 48% of those denied were truly disabled (type I errors). Low and Pistaferri (2015) estimates these errors using a structural model of the DI application process and find type II errors ranging up to 18%, depending on age, and type I errors are 37% for those age 45, and higher for those under 45. Similar numbers are found by Low and Pistaferri (2019) using merged administrative and survey data, with the type II errors of 28% and type I errors of 54%. If individuals recover but do not flow off DI, we would expect the fraction falsely claiming to be higher in the stock than at admission. This is the finding of Benitez-Silva, Buchinsky, and Rust (2004) who use self-reported disability data on the over 50s from the Health and Retirement Study (HRS): over 20% of recipients of DI are not truly work limited.

How important these numbers are depends not just on the size of the errors but also on the numbers of healthy and disabled applying. To the extent that there is already a lot of self-selection of the disabled into applying, the numbers of falsely rejected are even more concerning. The uncontrovertible aspect of these numbers however is that clearly it is not straightforward for the social security administration to assess disability accurately.

Incentives to apply for DI will be affected by poor labor market conditions, such as declines in individual productivity due to negative shocks to skill prices or low arrival rates of job offers. Some papers have used aggregate economic shocks to study participation in disability insurance programs. Black, Daniel, and Sanders (2002) study the impact of the boom and bust in the coal mining industry of the 1970s and 1980s. Coal prices increased in the 1970s as a consequence of the oil shocks; they then declined sharply in the 1980s and 1990s when oil prices stabilized. Areas rich with coal in the US (such as the Appalachian region, containing parts of Kentucky, Ohio, Pennsylvania, and West Virginia) were dramatically affected by these events, with employment from mining booming during the 1970s and then collapsing when coal prices declined. Black et al. (2002) use variation in local earnings growth within states (which they argue represent long-term changes, rather than transitory ones) to test whether the exogenous changes in the local economy's fortunes induced by the boom and bust of the coal mining industry affected participation in the DI program (measured as local spending on DI and SSI). They found sizable elasticities of DI and SSI utilization with respect to local earnings changes (0.35 and 0.55, respectively). Further, Black et al. (2002) show that participation in the DI program is much more likely for permanent than transitory skill shocks. Benítez-Silva, Disney, and Jiménez-Martín (2010) use cross-country micro data and relate disability claims to local unemployment rates, controlling for subjective measures of health status. Interestingly, they document that the increase in DI claims appears to be caused by fewer people exiting disability insurance programs rather than more people entering them.

Comparison with smaller UI take-up rates confirms that elasticities are larger in response to permanent job creation or destruction (as those induced by the bust of the coal mining industry) than in response to more temporary shocks in the local labor market.

Some recent papers have extended this idea to look at the impact of the 2001 recession and of the great recession (Lindner, Clark, and Javier (2017), Maestas, Mullen, and Strand (2015). In particular, Maestas, Mullen, and Strand (2015) show that during the Great Recession DI applications increased by 1.3 percent for each percentage point increase in the unemployment rate; however, the award rate did not change significantly. This suggests most of the "induced" claims come from marginal applicants.

Mullen and Staubli (2016) use Austrian administrative data to analyze the effect of benefit generosity on DI claims and applications. For identification, they exploit exogenous variation in benefits arising from several reforms to the Austrian DI and old age pension system that took place in the 1990s and 2000s. They find that in the 2004-10 period, the elasticity of DI claiming with respect to DI benefit generosity is 0.7 while the elasticity of applications with respect to benefits is 1.6. This implies that application screening considerably mediates the effect of increased DI generorsity leading to increased numbers of new DI beneficiaries.

de Jong, Lindeboom, and van der Klaauw (2011) use a controlled experiment conducted in the Netherlands to look more directly at the effect of more stringent screening of applicants. In 2003, caseworkers at the National Social Insurance Institute (which is the institution responsible for the screening of disability insurance applicants) were instructed to screen applicants more strictly than usual. The experiment was conducted in two "treatment" regions, leaving other regions as controls. Using simple diff-in-diff analysis, the authors find that stricter screening reduced DI applications, an effect they argue can be explained by improved self-screening by potential DI applicants and an increase in return-to-work activities during sickness absence.

4.1.3 Spousal Labour Supply, Savings and Other Insurance

Besides the impact on an individual's own labor supply, the DI program may in principle affect the labor supply of other household members (added worker effects). Further, changes in disability insurance generosity may affect the likelihood of applying to other welfare programs. Autor, Kostol, Mogstad, and Setzler (2017) use Norwegian administrative data and a "judge assignment" research design, to show that DI increases household income and consumption significantly for single-person hosueholds. They also find that DI receipt reduces usage of other transfer programs, but by less than one-to-one, so that total benefits increase. However, married households on average do not see an increase in income and consumption, as the recipient's spouse adjusts their labor decisions in response to DI decisions.

Similarly, **?** investigates the effect of cessation in SSI payments to children on their parents' labor effort, using a regression discontinuity design based upon cuts in budget for medical evaluations of children on SSI in 2005, which reduced the chance of being dropped from the program. **?** finds that parents respond to being dropped from the program by increasing labor effort, approximately offsetting the children's lost disability income one-for-one. Futhermore, they do not generally substitute towards other transfer programs.

Mueller, Rothstein, and von Wachter (2016) study the interactions between the UI and the DI program after the Great Recession, showing that DI applications are countercyclical. During the Great Recession the UI program was extended considerably from the statutory 26 weeks to up to 99 weeks. In principle, this would have slowed down applications to the DI program if the two programs are substitutes, and then increased applications to DI once people start exhausting their UI benefits. In practice, the authors find small to negligible effects because there is very little overlap in the population of UI and DI recipients.

4.1.4 Reassessment and the Return to Work

Less studied is the effect of policies that create incentives for current DI beneficiaries to return to work, even though in principle the response should depend on the same parameters that govern the decision to transition from work into DI. Kostol and Mogstad (2014) study a return-to-work policy that was introduced in Norway in 2005. In both the US and Norway, DI beneficiaries face a 100% tax rate if they work and earn above a certain threshold (the "substantial gainful amount", or SGA). The program studied by Kostol and Mogstad (2014) reduced the penalty for earning above the SGA; moreover, individuals awarded DI before 2004 were penalized less harshly for earning above the SGA compared to those whose awards came after 2004. This created a discontinuity that the authors use for identification. The authors find that the policy significantly increased labor force participation and earnings for recipients, did not reduce their welfare, and reduced program costs. This suggests that many DI recipients actually have significant capacity to work.

In the US, the “Ticket-to-work” program was a policy that tried to get DI beneficiaries back to work by offering free employment services, but estimates of its effects were negligible (see Thornton et al. 2006). Various states have experimented with so-called “\$1 for \$2” pilots, in which, similar to the program studied by Kostol and Mogstad (2014), DI beneficiaries who work above the SGA face a reduced implicit tax rate up to some amount (a tax rate of 50% instead of 100%), and can keep Medicare benefits, up to some time limit. Benitez-Silva et al. (2005) use the HRS and focus on older workers to study the “\$1 for \$2 benefit offset”. They estimate only a very small effect of the reform on returning to work. Their model is very detailed in numerous dimensions, but one important caveat is that there is no disaggregation of the response to these incentives by the severity of health status. As stressed by von Wachter et al. (2011) , behavioral responses to incentives in the DI program differ by age and by health status, with the young being the most responsive. There is an ongoing large scale experiment involving six states (the Benefit Offset National Demonstration, or BOND) that implements the same “\$1 for \$2” format, but no results have been published at the time of writing.

In response to a rapid increase in DI participation over the 1980s and early 1990s, in 1993 the Dutch government passed a major reform aimed at checking more thoroughly the work capacity of young DI recipients (below age 50). Moreover, since the Dutch system allows for partial disability, the more stringent criteria on assessing work capacity induced a decline in benefits for those who remained on the program. Exogenous variation in disincentives to stay on the program comes partly from the age criterion, and partly from the fact that the program was phased in by year-of-birth cohort. Borghans, Gielen, and Luttmer (2014) exploit this reform to look at how return to work is affected by changes in incentives to stay on the program. Using a regression discontinuity design, they find that the stricter re-assessment rules reduced benefits on average by 1100 euro per year. A small fraction of this decline (about 30%) was attributable to people leaving the program, while the bulk was due to a decline in benefits for those who remained on the program. Next, the authors ask if the decline in benefits was replaced by increased employment/earnings, increased use of other social insurance programs, or whether it was absorbed by a decline in consumption (living standards). They conclude that disability recipients were able to offset almost all of the reduction in benefits induced by the reform by either increasing employment and earnings (around 2/3) or by more intense use of other welfare programs (the remaining 1/3), with little effect on the work of other family members, etc. This supports the view that those who got cut out of the program because of more stringent re-assesment rules had substantial remaining work capacity.

Bianco (2019) uses a structural life-cycle framework in the UK environment to simulate the consequences of employment subsidies for disability recipients going back to work. Her central conclusion is that a significant number of individuals on DI have work capacity and are responsive to incentives to return to work, with about 30% of recipients taking up work.

On the other hand, the evidence from the US, is that following the mass removals in 1996, the vast majority had moved back onto DI by different routes within a few years. reference

4.2 Value of Disability Insurance

A narrow approach to considering the value of insurance is to consider the consumption loss associated with a negative health shock. In the US, the consumption loss is considered by Meyer and Mok (2018) and also by Gallipoli and Turner (2009). In the UK, this is considered by Ball and Low (2014). Consumption in both the UK and the US falls on disability, by about 12% for the US, and 9% for the UK. to double check These falls are mitigated by the receipt of disability insurance. However, as discussed in Ball and Low (2014), there is selection into who actually receives disability insurance such that those hit by the worst shocks are more likely to be in receipt of DI and more likely to face large consumption losses. This implies that estimates are lower bounds of the true mitigation achieved by DI. Meyer and Mok (2018) use these estimates of consumption loss in a Bailey-Chetty framework to consider the optimality of making DI more generous. What this approach cannot capture is the importance of the persistence of shocks, and how the details of the DI scheme, such as the screening process, really matter.

The broader issue of the value of DI and the effects of DI reform requires knowing preferences, constraints and risk of health and other shocks and how these evolve across individuals' lifetimes. These can be used to calculate expected utility at the start of life. We could then evaluate how expected utility is affected by disability insurance programs that mitigate the risk, accounting for the fiscal cost of the program and for the changes in labour supply and saving behaviour resulting from the program. These calculations clearly involve a cardinalisation of preferences and introducing substantial structure, as discussed in section 3.

Work by Bound, Cullen, Nichols, and Schmidt (2004) and Bound, Stinebrickner, and Waidmann (2010), Benitez-Silva, Buchinsky, and Rust (2005), Waidman, Bound, and Nichols (2003) and Low and Pistaferri (2015) has highlighted the importance of considering both sides of the insurance/incentive trade-off for welfare analysis and conducted some policy experiments evaluating the consequences of reforming the program. These papers differ in focus and this leads to differences in the way preferences, risk, and the screening process are modeled; and in the data and estimation procedure used.[15]

Low and Pistaferri (2015) stress that we need a life-cycle perspective to capture fully the insurance benefits, and we need an accurate characterization both of labor supply behavior and applications to the program to capture fully the incentive costs of the program. This perspective leads to three key conclusions about the US system: first, individuals have very little capacity to self-insure disability shocks, in contrast to their ability to self-insure unemployment shocks, which are much more transitory. Second, there are substantial false rejections in the disability insurance process, leading many individuals who are in need of support to rely on a very minimal welfare state rather than receiving DI. Other individuals are discouraged from applying. By contrast, false positives are much less prevalent: this incentive cost of the DI program is not a first-order issue. Finally, the labour supply of those with disabilities would be very low even in the absence of the DI program and so these incentive costs are also muted. On the other hand, there are labour supply distortions for those with moderate disabilities who may have applied because of low productivity. This raises the broader question of whether the role of government is about providing insurance against low standards of living for whatever reason, or whether there is something specific about disability. They calculate the welfare implications of various reforms by measuring the willingness to pay (in consumption terms) for the new policy, i.e., the fraction of consumption that would makes individual indifferent ex ante (behind the veil of ignorance) between the status quo and the policy change considered. All experiments are government-budget neutral, although there are no general equilibrium effects.

Deshpande, Gross, and Su (2019) look at the impact of outcomes of disability insurance application on financial outcomes, such as the likelihood of declaring personal bankruptcy, etc.

Amanda Michaud's work

5 Policy Implications

Given the accumulating evidence on the distortions induced by increasing generosity of disability insurance (as well as the undeniable importance of insurance for those who are genuinely hit by the disability shocks), we conclude this survey with a discussion of the policy implications of potential

[15] There is a purely theoretical literature on optimal disability insurance, such as the model of Diamond and Sheshinski (1995) and the model of Golosov and Tsyvinski (2006) on the desirability of asset testing DI benefits.

reforms. We consider the evidence in each of the 5 policy dimensions outlined in section 2.

Medical Tests and Stringency

A first reform to consider is changing the strictness of the screening process. In the US a reform like this was implemented in the early 1980s and led to sharp declines in inflows onto DI and significant removal of DI recipients, although the criteria were relaxed again in 1984 after a political backlash. Gruber and Kubik (1997) study cross-sectional heterogeneity in strictness across DDS centers. Low and Pistaferri (2015) find that increasing the stringency of the screening process by raising the disability threshold for admission into the programs reduces the extent of the incentive problem, but also reduces the extent of insurance provided by the program as expected. This reduces welfare (expected utility) and is a clear example of the trade off of incentives and insurance.[16] Part of the reason for this conclusion that reduced strictness is welfare increasing is the low acceptance rate of severely disabled individuals onto DI. The subgroup of young severely-disabled individuals are particularly ill-equipped to insure against disability risk because these individuals face high rejection rates when applying for DI and yet have not had time to accumulate enough assets to self-insure. Hence reduced strictness that increases the chance to get into the program is highly valued.

The Application Process and Labour Force Attachment

A clear contrast between the US and other countries is the attempt to maintain attachment of applicants to the labour force. In the Netherlands, this has been achieved partly through the closer involvement of firms de Jong et al. (2011), and in the UK, through providing incentives for job centres to support applicants (Petronoglo and Van Reenen, 2019).

Eligibility Requirements

The DI program may interact in important ways with other government welfare programs, such as food stamps or income support. Low and Pistaferri (2015) investigate the importance of such interactions by changing the generosity of the means-tested program. Marginal applicants to DI switch their decision from applying to not-applying: for ”false applicants”, the means-tested program acts as a *substitute* for DI and generally applications to DI fall as income support generosity increases.

[16] 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.

This is because at some point income support provides such a sufficiently generous support, and without the uncertainty of applications for DI, that false applications for DI fall. By contrast, for severely disabled workers income support is *complementary* to DI: the fraction of the severely disabled who receive DI increases as income support become more generous. This is because the consumption floor increases, making application for DI less costly for the severely disabled who were marginal between working and applying for DI. The effect of increasing food stamps generosity is therefore welfare improving, as it reduces the extent of false applications while increasing insurance for the truly disabled at the margin, as well as providing insurance for those whose skills have become obsolete. Part of the reason for this result is that the income support is less distortionary than DI because it does not require people to disengage from the labor force and to stop working altogether.

Generosity and Progressivity

One of the lessons from the trade-offs captured inLow and Pistaferri (2015) is that increases in DI generosity can be welfare improving despite the increase in moral hazard (false applications) it generates.[17] The point is that the greater insurance value of more generous payments dominates the cost of the revenue needed to pay the false claimants (in the form of higher tax rate on workers), although the effect varies substantialy with the underlying productivity of individuals.

Reassessment and the Return to Work

Some US commentators have pointed out that there are cases in which individuals with mild to moderate disabilities value the medical care they receive under the DI program more than the cash benefits component of the program. In the pre-ACA era, this may have been particularly relevant for individuals without access to health insurance due to pre-existing conditions. There is little evidence for the importance of this mechanism. The Ticket-to-Work experiment allowed DI beneficiaries who were transitioning to work to retain their Medicare benefits for up to 7 years follwoing their 9-month Trial Work Period. However, take up rates for this program have been substantially low. There is no research to date on the impact of ACA (which eliminated denials of private health insurance because of existing pre-conditions) on DI application rates. On the other hand, we have reported evidence that those on disability have residual work capabilities.

[17] Meyer and Mok (2018) reach a similar conclusion. They apply a variant of the benefit optimality formula derived by Chetty (2008) to conclude that the current level of DI benefits is lower than the optimal level and that it‘ is welfare improving to increase DI generosity.

Last Thoughts

In terms of the accuracy of the DI programs, the main lesson is the substantial false rejections of those who are in need of insurance. Only 58% of the severely work-limited are in receipt of DI. By contrast, the number of false applications appears much less serious. Similarly, while there is evidence of labour supply disincentives induced by DI, these are not large: neither recipients of DI nor those rejected from DI participate much in the labour force. On the other hand, there is evidence that it is very difficult to incentivise or move people off DI once they are on it, whether because of a lack of labour market attachment, skill depreciation or individual types.

The final policy conclusions to draw from this survey are that less stringent testing of the program, coupled with labour market rehabilitation from the moment of application, are more effective ways of providing effective insurance for those in need alongside minimising the extent of false recipients. There are no simple solutions to the incentive-insurance trade-off. However, creative rather program design can reduce incentive costs and improve insurance, as shown by the Netherlands through the 2000s.

References

Autor, D., A. R. Kostol, M. Mogstad, and B. Setzler (2017, June). Disability benefits, consumption insurance, and household labor supply. Working Paper 23466, National Bureau of Economic Research.

Autor, D. H. and M. G. Duggan (2006, September). The growth in the social security disability rolls: A fiscal crisis unfolding. *Journal of Economic Perspectives 20*(3), 71–96.

Ball, S. and H. Low (2014, July). Do Self-insurance and Disability Insurance Prevent Consumption Loss on Disability? *Economica 81*(323), 468–490.

Banks, J., R. Blundell, and C. Emmerson (2015, May). Disability benefit receipt and reform: Reconciling trends in the united kingdom. *Journal of Economic Perspectives 29*(2), 173–90.

Benitez-Silva, H., M. Buchinsky, H. M. Chan, J. Rust, and S. Sheidvasser (1999, June). An empirical analysis of the social security disability application, appeal, and award process. *Labour Economics 6*(2), 147–178.

Benitez-Silva, H., M. Buchinsky, and J. Rust (2004, January). How Large are the Classification

Errors in the Social Security Disability Award Process? NBER Working Papers 10219, National Bureau of Economic Research, Inc.

Benitez-Silva, H., M. Buchinsky, and J. Rust (2005, March). Induced Entry Effects of a $1 for 2$ Offset in SSDI Benefits. Department of Economics Working Papers 05-03, Stony Brook University, Department of Economics.

Benítez-Silva, H., R. Disney, and S. Jiménez-Martín (2010, July). Disability, capacity for work and the business cycle: an international perspective. *Economic Policy 25*, 483–536.

Bianco, C. D. (2019). Disability insurance and the effects of return-to-work policies. *University of Padua, mimeo*.

Black, D., K. Daniel, and S. Sanders (2002, March). The impact of economic conditions on participation in disability programs: Evidence from the coal boom and bust. *American Economic Review 92*(1), 27–50.

Borghans, L., A. C. Gielen, and E. F. P. Luttmer (2014, November). Social support substitution and the earnings rebound: Evidence from a regression discontinuity in disability insurance reform. *American Economic Journal: Economic Policy 6*(4), 34–70.

Bound, J. (1989). The health and earnings of rejected disability insurance applicants. *The American Economic Review 79*(3), 482–503.

Bound, J. and R. V. Burkhauser (1999, May). Economic analysis of transfer programs targeted on people with disabilities. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3 of *Handbook of Labor Economics*, Chapter 51, pp. 3417–3528. Elsevier.

Bound, J., J. B. Cullen, A. Nichols, and L. Schmidt (2004, December). The welfare implications of increasing disability insurance benefit generosity. *Journal of Public Economics 88*(12), 2487–2514.

Bound, J., T. Stinebrickner, and T. Waidmann (2010, May). Health, economic resources and the work decisions of older men. *Journal of Econometrics 156*(1), 106–129.

Burkhauser, R. and M. Daly (1996, 08). Employment and economic well-being following the onset of a disability: The role for public policy.

Burkhauser, R. and M. Daly (2011, 01). The declining work and welfare of working-age people with disabilities.

Chen, S. and W. van der Klaauw (2008, February). The work disincentive effects of the disability insurance program in the 1990s. *Journal of Econometrics 142*(2), 757–784.

Chetty, R. (2008, April). Moral Hazard versus Liquidity and Optimal Unemployment Insurance. *Journal of Political Economy 116*(2), 173–234.

de Jong, P., M. Lindeboom, and B. van der Klaauw (2011). Screening disability insurance applications. *Journal of the European Economic Association 9*(1), 106–129.

Deshpande, M., T. Gross, and Y. Su (2019, March). Disability and Distress: The Effect of Disability Programs on Financial Outcomes. NBER Working Papers 25642, National Bureau of Economic Research, Inc.

Diamond, P. and E. Sheshinski (1995). Economic aspects of optimal disability benefits. *Journal of Public Economics 57*(1), 1–23.

French, E. and J. Song (2014, May). The effect of disability insurance receipt on labor supply. *American Economic Journal: Economic Policy 6*(2), 291–337.

Gallipoli, G. and L. Turner (2009, June). Household Responses to Individual Shocks: Disability and Labour Supply.

Golosov, M. and A. Tsyvinski (2006, April). Designing optimal disability insurance: A case for asset testing. *Journal of Political Economy 114*(2), 257–279.

Grossman, M. (1972). *The Demand for Health: A Theoretical and Empirical Investigation.* Columbia University Press.

Gruber, J. and J. D. Kubik (1997). Disability insurance rejection rates and the labor supply of older workers. *Journal of Public Economics 64*(1), 1–23.

Kostol, A. R. and M. Mogstad (2014, February). How financial incentives induce disability insurance recipients to return to work. *American Economic Review 104*(2), 624–55.

Kreider, B. (1999). Latent work disability and reporting bias. *Journal of Human Resources 34*(4), 734–769.

Lillard, L. and Y. Weiss (1997). Uncertain health and survival: Effects on end-of-life consumption. *Journal of Business and Economic Statistics 15*(2), 254–68.

Lindner, S., B. Clark, and M. Javier (2017). Characteristics and employment of applicants for social security disability insurance over the business cycle. *The B.E. Journal of Economic Analysis and Policy 17*(1).

Low, H. and L. Pistaferri (2015). Disability insurance and the dynamics of the incentive insurance trade-off. *American Economic Review 105*, 2986–3029.

Low, H. and L. Pistaferri (2019). Disability insurance and gender differences: Evidence from merged survey-administrative data. *University of Oxford mimeo*.

Maestas, N., K. J. Mullen, and A. Strand (2013, August). Does disability insurance receipt discourage work? using examiner assignment to estimate causal effects of ssdi receipt. *American Economic Review 103*(5), 1797–1829.

Maestas, N., K. J. Mullen, and A. Strand (2015, May). Disability insurance and the great recession. *American Economic Review 105*(5), 177–82.

Meyer, B. D. and W. K. Mok (2018). Disability, earnings, income and consumption. *Journal of Public Economics*.

Mueller, A. I., J. Rothstein, and T. M. von Wachter (2016). Unemployment insurance and disability insurance in the great recession. *Journal of Labor Economics 34*(S1), S445–S475.

Mullen, K. J. and S. Staubli (2016). Disability benefit generosity and labor force withdrawal. *Journal of Public Economics 143*(C), 49–63.

Nagi, S. (1969). *Disability and Rehabilitation.* Ohio State University Press.

Parsons, D. O. (1980). The decline in male labor force participation. *Journal of Political Economy 88*(1), 117–134.

von Wachter, T., J. Song, and J. Manchester (2011, December). Trends in employment and earnings of allowed and rejected applicants to the social security disability insurance program. *American Economic Review 101*(7), 3308–29.

Waidman, T., J. Bound, and A. Nichols (2003, April). 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.

Wu, A. Y. and J. S. Hyde (2018). The postretirement well-being of workers with disabilities. *Journal of Disability Policy Studies 0*(0), 1044207318793161.