FISCAL STUDIES, vol. 41, no. 1, pp. 129–164 (2020) 0143-5671

Disability Insurance: Theoretical Trade-Offs and Empirical Evidence*

HAMISH LOW† and LUIGI PISTAFERRI‡

†*University of Oxford; Institute for Fiscal Studies*
(hamish.low@economics.ox.ac.uk)
‡*Stanford University; SIEPR; NBER; CEPR*
(pista@stanford.edu)

Abstract

Disability insurance provides protection against health shocks that limit the ability to work. In most countries, these programmes are large and growing, both in expenditure and in number of recipients. We discuss the traditional trade-off between insurance and incentives in providing this insurance, with a focus on the US and UK experiences. There is substantial evidence on the extent of the 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. We provide a life-cycle framework for understanding the trade-offs and to evaluate the welfare implications of policy reforms. We argue that reforms should be focused on reducing false rejections and supporting labour market attachment. The difficulty in considering reform is that the design of disability insurance has many aspects that interact and impact on outcomes.

* Submitted April 2019.
Thanks to two anonymous referees for suggestions, to Tom Waters for help with the data and to Max Rong for research assistance.
Keywords: social insurance, incentives, false rejections.
JEL classification numbers: H53, J65, H55.

I. 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 programmes has been growing fast, measured both in terms of total spending and in terms of the fraction of the population in receipt of benefits. Compared with the analysis of unemployment insurance (UI), the economics literature has devoted less attention to disability insurance (DI) programmes, although this situation is changing fast. Much of the current debate about the scale of the programmes focuses on the incentive costs, particularly in terms of reduced labour supply. On the other hand, and less researched, these programmes have a substantial value for 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 on the 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 targeting, how to reduce false rejections and how to improve labour force attachment, rather than on how to reduce false applications per se.

Growth in disability insurance programmes can be explained partly by declines in health (at least for some demographic groups),[1] partly by changes in economic incentives and opportunity costs, and partly by changes in institutions. The open question is how much of this growth can be attributed to increasing recognition of mental and musculoskeletal disorders as work-related disabilities, and how much to acceptances onto the programme 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 discuss recent empirical evidence on the different sides of this trade-off and to provide a framework for thinking about its economic and policy implications.

Increased benefit generosity creates incentives for people to leave work and apply for the programme, to stay on the programme even when their health status improves and, in extreme cases, to exaggerate their disability in order to be awarded benefits. Suppose that individuals have a level of productivity that depends on health and skills. Some individuals with a moderate disability may choose to apply for DI if their skills deteriorate due to external shocks,

[1]Lakdawalla, Bhattacharya and Goldman (2004) cite the obesity crisis, especially affecting the low-income population, while Morden et al. (2014) show that the opioid crisis in the US, which increasingly affects the young, has contributed to increasing mortality rates in these groups after a secular decline and to a decline in exits from disability insurance rolls. Case and Deaton (2017) show declining patterns of health across birth cohorts for low-educated households at almost all ages of the working life.

such as automation or international trade, despite this not being the intended use for DI. The issue is that the 'true' disability status of an individual is unobserved and the screening imperfect. The screening process in DI is in fact much more difficult than that in UI, 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 arise because 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 of the reject/award type (as in the US) rather than deciding how fractionally disabled a person is (as in Italy or in the US veterans' disability programme). Another difficulty is that given the low exit rates from the programme, DI acts effectively as insurance against permanent shocks: it meets demand for protection against long-term unemployment or productivity declines that cannot be satisfied by other social insurance programmes (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 declines in productivity which in turn led to poor health.

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

The survey is structured as follows. We start in Section II with a discussion of institutional details about disability insurance programmes, focusing mainly on the US and the UK, and of the different dimensions of policy design, such as the nature of the medical test and the degree of progressivity. The trade-off between provision of insurance and the disincentives that this may create pervades all social insurance programmes. Section III discusses the empirical evidence on incentive effects and insurance implications. The focus in the empirical literature, perhaps because the issues are much better

[2]Benítez-Silva, Buchinsky and Rust (2004) have also emphasised 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.

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. Section IV provides a theoretical framework for discussing the insurance–incentives trade-off. Finally, Section V discusses policy implications.

II. Institutional detail and statistics

Many OECD countries offer support to those with health conditions that affect their ability to work, but the details of the support and the consequences for the wider economy vary enormously across countries. In this section, we contrast the scale of disability programmes over time and focus mainly on comparing the US and the UK, with a discussion of the experience and institutional aspects of other countries at the end of the section. We also discuss alternative aspects of the institutional structure, which determine the incentive and insurance effects of the programmes.

We use the broad term 'disability insurance' to cover programmes that offer income to replace earnings, where the loss of earnings is due to bad health. In the UK, this includes programmes such as invalidity benefit, incapacity benefit and employment & support allowance. In the US, this definition 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 earning capability due to poor health. By contrast, we use the term 'disability benefits' to cover programmes that offer direct help with costs of being disabled whether an individual is working or not. In the US, this includes coverage through Medicare or Medicaid of healthcare costs due to a disability. In the UK, this definition includes programmes such as disability living allowance, severe disablement allowance, attendance allowance and personal independence payments. We use the term 'disability programmes' to refer to the total of disability insurance and disability benefits. Our focus is mostly on disability insurance because of its size and the more direct link with labour market outcomes.

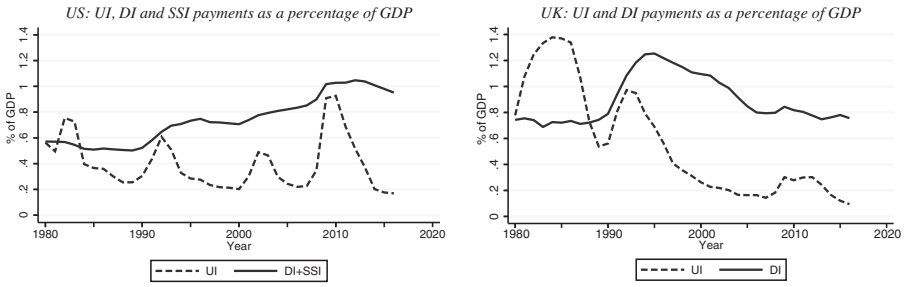
1. The scale of disability programmes

We can measure the scale of disability programmes through expenditure and through the number of recipients. Figure 1 shows how spending on disability insurance as a fraction of GDP has increased over the past 35 years in the US and in the UK. The figure also shows spending on UI. The first point to stress is the scale of the disability insurance programmes: approximately 1 per cent

[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.

FIGURE 1

Spending on disability and unemployment insurance

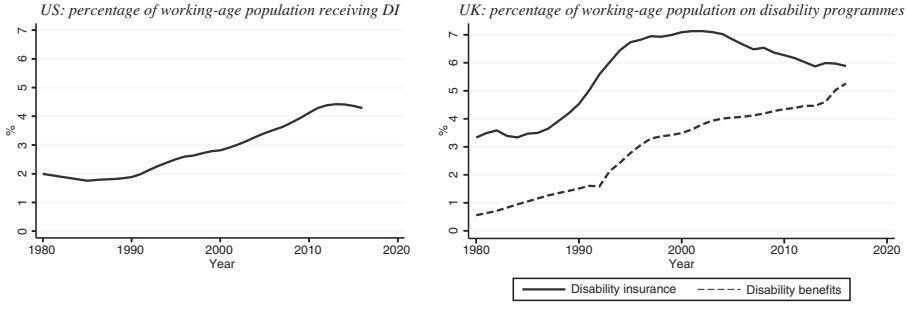


Note: Federal DI payments to disabled workers only. Federal SSI payments to the blind and disabled only.
Source: SSI 2018 Annual Report; Social Security Administration (SSA) 2017 Annual Statistical Supplement; Federal Reserve Economic Data (FRED).

Note: DI is total spending on all programmes classified as incapacity benefits for working-age people.
Source: Department for Work and Pensions (DWP) 2018 Autumn Budget Outturn and Forecast.

FIGURE 2

Number of beneficiaries of disability programmes



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

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

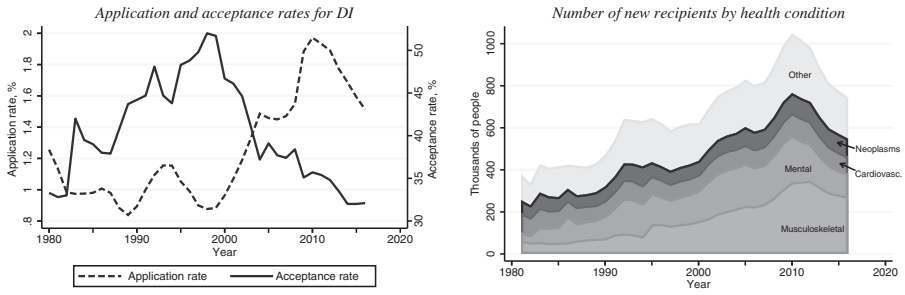
of GDP is now spent on disability insurance in the US, and slightly less in the UK. Spending on disability insurance is typically much larger than spending on UI.

Figure 2 shows the growth in recipiency rates for the US and the UK. In the US, 4.5 per cent of the working-age population receive DI, and the programme has been growing at a fairly constant rate since a reform in 1984 relaxed the admission criteria. In the UK in 2017, 6 per cent of the working-age population received disability insurance through the employment & support allowance, although this is below the peak of over 7 per cent of the early 2000s.[4] 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

[4]See also Banks, Blundell and Emmerson (2015).

FIGURE 3

Application rates and recipiency: US



Note: Application rate defined as number of applications divided by population insured in event of disability.
Source: SSA 2017 Annual Statistical Supplement.

Source: 2017 Annual Statistical Report on Social Security Disability Insurance (SSDI).

receiving disability benefits which directly support the extra financial costs of a disability. This rise is despite the reform of 2010 aimed explicitly at reducing this support.

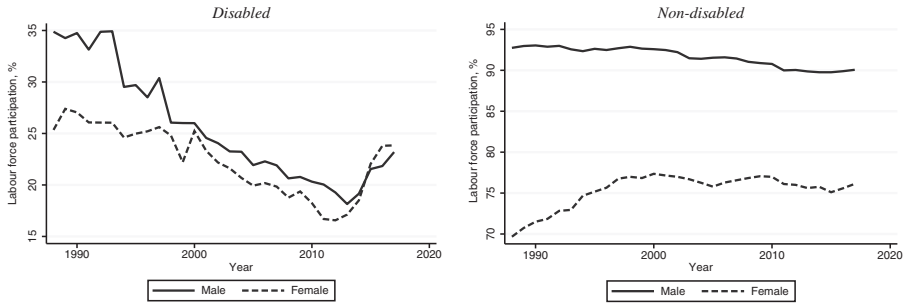
For the US, we are able to break these numbers of recipients down further. The left-hand graph of Figure 3 shows the evolution of the application and acceptance rates for DI: application rates have increased sharply since 2000, and with a corresponding fall in the acceptance rate. The net effect, however, is an ongoing increase in the number of new recipients, as shown in the right-hand graph. This graph also shows how the increase is split between different health conditions: the growth in the number of new recipients is among those with mental health conditions and musculoskeletal/back pain. This growth in new recipients translates into the large increases in the stock of recipients shown in Figure 2 because those entering with back pain or mental health problems tend to be younger and so stay on the programme for longer. Indeed, the fraction of new awards going to workers aged under 50 increased from 0.33 in 1980 to 0.41 in 2010.

Figure 4 shows what has happened to labour force participation in the US over the same time span using Current Population Survey (CPS) data. The striking, albeit unsurprising, point is that labour force participation rates are much lower for people with disabilities than for those who are not disabled.[5] As we discuss in the modelling of disability in Section IV, this lower 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 group. Finally, it may (perversely) reflect legislation such as the Americans

[5]In the CPS, all household members aged 18–64 are asked whether they have a health problem or disability that prevents them from working or that limits the kind or amount of work they can do. We classify as 'disabled' those who answer in the affirmative.

FIGURE 4

Labour force participation: US



Source: Current Population Survey, various years.

with Disabilities Act (ADA).[6] In addition to labour force participation being low, there has been a sharp decline in participation among those who are disabled. For men with a disability, participation rates have declined from 35 per cent to 23 per cent over the past three decades.

2. Dimensions of policy

This discussion of the evolution of disability insurance programmes and how the programmes differ between the US and the UK highlights the importance of understanding the dimensions of disability insurance policy. Programmes differ in five main dimensions:

- the nature of the medical test;
- the application process and labour market attachment during application;
- eligibility – whether the programme is contributory, requiring individuals to have made social-insurance-type payments before claiming, or means-tested;
- the generosity and progressivity of benefits;
- the process of reassessment and having benefits removed.

We discuss these dimensions in turn. As we show in Section III, 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 programmes to be capturing institutional details. Instead, the empirical evidence is on the total effect of a particular disability insurance system.

[6]See DeLeire (2000) and Acemoglu and Angrist (2001).

a) 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, worsening insurance.

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 feasible, the applicant would still receive disability insurance. This more lenient definition is related to the notion of disability as being a loss of earning 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 are taken into account, the absence of such an 'appropriate' or 'suitable' job might be the factor that puts the individual into the disabled group. Similarly, if the criterion is a loss of earning capacity, then the disability definition may be tied to 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 earning capability but not prevent them from working altogether.

These broader economic definitions of disability lead to an assessment of more individuals as being disabled and to the numbers going onto disability insurance varying with the business cycle.[7] This in turn blurs the distinction between unemployment insurance and disability insurance.

The breadth of definition used to determine disability insurance eligibility varies across countries. The UK's approach is to concentrate on what an individual is still able to do following a health shock, using residual work capacity. In contrast, the US defines those unable to earn more than a certain monthly threshold due to a health condition as being eligible.

There are two further crucial issues concerning medical evidence. The first concerns who actually reviews the evidence and makes the decision – 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 its own doctors. An intermediate case is where the individual's own doctor submits evidence to the administration, which is then reviewed 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 mental illness, musculoskeletal pain, drug or alcohol addiction, and so on.

[7]Black, Daniel and Sanders, 2002; Benítez-Silva, Disney and Jiménez-Martín, 2010.

In the US, individuals submit written evidence from their own doctors or medical providers which is assessed by the Social Security Administration (SSA) local field 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.[8] 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 per cent to 33.3 per cent between 1975 and 1980, but then rose again quickly in 1983, around the time that eligibility criteria were liberalised and an applicant's own physician reports were used to determine eligibility, to a peak of over 50 per cent in 1998. Further, although in 1983 82 per cent of initial DI awards were made strictly on medical criteria, by 2001 this had dropped to just 58 per cent.[9]

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 his or 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 put them into the non-disabled group. 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. In this way, individuals are divided into the 'work-related activity group' and 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.

b) 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 assessment; and second, the degree of labour market detachment during the application process.

The US DI programme is very much targeted at individuals who are out of the labour market, and labour market detachment is required. There is a

[8] Awards are made either for strictly medical reasons (listed impairments that are easily classifiable) or for vocational reasons (inability to perform previous jobs or jobs that befit one's skills, age, etc., because of a medical condition).

[9] Autor and Duggan, 2006.

statutory five-month waiting period out of the labour force from the onset of disability before an application for DI can be filed. For those who are initially rejected, further appeals and processing mean the average time until the final acceptance decision is substantial.[10] For 2006, the SSA reported that the (cumulative) time elapsing between filing a DI application and obtaining a decision was 131 days for the initial decision, 279 days for a reconsideration and 811 days for a decision on appeal by administrative law judges.

Perhaps one 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 systems in 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 insurance have to have alternative means of living or rely on means-tested benefits. By contrast, in 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 – for example, the US, which offers no vocational rehabilitation when health shocks strike, in addition to requiring detachment from the labour force in order to apply. The recent move in the UK to identify a work-capability group was intended to maintain engagement with working.

c) 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 are typically used and so the distinction determines primarily the amount of benefits received.

In the US, DI is the contributory programme that provides cash and healthcare benefits for covered workers, their spouses and dependants.[11] By contrast, SSI is the means-tested welfare programme: applicants must meet the same medical criteria as DI applicants but also have limited income and limited resources. The definition of low income and low resources is similar

[10] Benítez-Silva et al., 1999; French and Song, 2014.

[11] 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 five of the last ten 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 per cent.

to the one used for other federal welfare programmes, such as the food stamps programme.[12]

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 mixed system of contributory and non-contributory payments was changed with the introduction of employment & support allowance, which explicitly had separate contributory and non-contributory (income-based) components.

The strictness of the means test also varies substantially: in the UK, the asset test for income-based ESA is maximum savings of £16,000, which is substantially higher than the $2,000 of savings for individuals applying for SSI in the US.

d) Generosity and progressivity

Generosity varies substantially across countries. The US offers earnings-related benefits (for the DI programme), whereas the UK offers 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 formula used to compute social security retirement benefits.[13] While benefits are independent of the extent of the work limitation, caps on the payroll tax financing the DI programme 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 cash payments, individuals who receive DI also have their healthcare costs covered through the Medicare programme (what we called 'disability benefits' above). However, this coverage is only received after being on DI for two years. This provision of healthcare, albeit delayed, means the replacement rates are substantially higher for workers with low earnings and those without employer-provided health insurance.

The UK in 1995 reduced the generosity of cash benefits, by removing the component of benefits relating to past earnings. This change had a larger impact on those with long earnings histories. By contrast to the US, the UK disability insurance programme is now flat rate, although that flat rate differs according

[12] In particular, individuals must have income below a 'countable 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).

[13] 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). The formula for computing benefits 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).

to the disability group. Those entering the work-related activity group after April 2017 receive £317 a month, whereas those in the support group receive £484. Those eligible for the income-related component receive payments on top of this.

e) Reassessment process

There are two main determinants of individuals leaving DI, apart from exiting into retirement or death: (1) through reassessments and (2) through incentives to return to work.

In both the US and the UK, reassessments depend on the qualifying condition. In practice, reassessment is very infrequent, conducted in unreliable ways (for example, by mail) for cost reasons, and benefit awards are essentially permanent.

In the US, disability insurance beneficiaries have their disability reassessed periodically through Continuing Disability Reviews (CDRs). By law, the SSA is expected to perform CDRs every seven years for individuals where medical improvement is not expected, every three years for individuals where medical improvement is possible, and every six to eighteen 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 between 1981 and 1983 when there were 950,000 CDRs. Of these, 40 per cent resulted in a termination of benefits.[14] 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.[15] An attempt in 1996 to reduce the number of disability insurance claimants by narrowing the criteria to exclude alcoholism and drug addiction led to many of those excluded returning onto benefits under a different disability.[16]

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

[14] At the same time, admission criteria were made stricter and the payroll tax raised, which puts emphasis on the fact that most policies touch various aspects of behaviour at once rather than being implemented in isolation, making policy evaluations complicated. In Section III, we discuss some quasi-experimental evidence that attempts to focus on the effects of 'cleaner' policy changes.

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

[16] Rupp and Stapleton, 1998.

limitations and so allocated to the support group, there is no reassessment. By contrast, individuals allocated to the work-related activity group are expected to improve and are asked to go to work-focused interviews to increase their chances of working in the future. Further, this work-related group will be allocated a referral date when 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 the US, for those in receipt of disability insurance, after nine months of earning more than the fairly low 'substantial gainful activity' threshold, benefits are withdrawn. Further, in 1999, a number of work incentive programmes for DI beneficiaries were introduced, such as the Ticket to Work programme, which allowed recipients to keep their Medicare eligibility for a number of years despite returning to work. The success of the programme has been limited.

3. Other countries

Most disability insurance programmes in OECD countries have institutional features that resemble those described above for the US and the UK. Most countries have also experienced a large increase in inflows and have debated reforms. Indeed, the fraction of working-age individuals receiving disability insurance is often higher in these countries than in the US and the UK (in 2013, the fraction was 7.4 per cent in the Netherlands and almost 10 per cent in Norway).

While the actual details differ, many countries condition eligibility to the presence of a medical criterion that impacts usual work and the absence of some residual functional capacity. Unlike the US, but similar to the UK, in most European countries disability insurance is the final stage of a process that starts with a worker drawing sickness benefits, which may be seen as playing the same role as the five-month waiting period in the US system (i.e. weeding out applications from people with temporary disabilities).

Unlike the US, a number of European countries first have potential disability insurance applicants undergo rehabilitation programmes before considering them for benefits. For example, in Denmark an individual must go through vocational rehabilitation before becoming eligible for benefits (apart from extreme cases), while in Sweden individuals can combine vocational rehabilitation with partial disability insurance. In the Netherlands, maintaining attachment of applicants to the labour force has been achieved partly through the closer involvement of firms.[17]

[17] De Jong, Lindeboom and van der Klaauw, 2011.

In terms of benefit structure, perhaps the most notable difference between the US/UK and other European countries is that in the former eligibility is a binary event, while in the latter applicants can be fractionally disabled and receive benefits correspondingly.

Countries differ significantly as regards the provision of incentives to return to work once on the programme and whether and for how long benefits can be kept while doing some paid work. 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.

III. Empirical evidence on the insurance–incentive trade-off

Health shocks that stop an individual being able to work are rare, but can be extremely serious: insurance should be very valuable. However, the fact that disability is a low-probability event makes self-insurance extremely costly. The private disability insurance market is plagued by traditional asymmetric information issues (i.e. unobservability of type). Government involvement in insuring individuals against health shocks arises precisely because of these market failures. While governments can force all workers to contribute to the financing of disability insurance through payroll or general taxation (thus eliminating adverse selection problems), this does not solve the insurance–incentive trade-off. In this section, we discuss the empirical literature that has studied this trade-off. In the next section, we discuss a general theoretical framework to capture the different issues highlighted in the empirical literature and which can be used to evaluate reforms.

1. Incentive effects of disability insurance

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 programme on labour 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 insurance–incentive trade-off. Below, we summarise 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. We cite evidence for other countries whenever available and appropriate.

a) Labour supply

In terms of estimating labour supply effects, the incentive for individuals to apply for DI rather than to work has been assessed by asking how many DI recipients would be in the labour force in the absence of the programme. The key difficulty is identifying an appropriate control group[18] 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 labour market outcomes for three groups of 45- to 64-year-old individuals observed at least 18 months after the initial application: DI beneficiaries; rejected DI applicants; and non-applicants. He finds that DI beneficiaries have very low employment rates (3 per cent), compared with rejected applicants (around 30 per cent) and non-applicants (75 per cent). Von Wachter, Song and Manchester (2011) replicate the approach of Bound (1989) with more recent administrative data. They compare labour market outcomes for four groups of individuals two years after their initial application, separating out individuals by the level at which the 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.[19] 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 per cent, as opposed to 25 per cent among those allowed at a later stage, 53 per cent among rejected applicants and 82 per cent among non-applicants. The difference between those rejected and accepted is around 30 percentage points, as in Bound (1989). Von Wachter et al. 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 analysed. Indeed, a key change in the composition of claimants since 1984 has been the growth in younger claimants.

The heart of this strategy is a comparison between those accepted onto the programme 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 hence rejected applicants' employment rates are an upper bound of how many DI beneficiaries would be working in the absence of the programme. In principle, one could address this issue by randomising awards and denials into the DI programme, an experiment that is of course infeasible.

[18] See Parsons (1980) and Bound (1989).

[19] 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).

There have been various attempts in the literature 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 (2019) extend the analysis of Bound (1989) and compare the post-application labour outcomes of awarded applicants and applicants who were rejected at different stages of the evaluation process. Finally, Maestas, Mullen and Strand (2013) and French and Song (2014) 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 find that those who had their appeals for disability insurance allowed by administrative law judges had a 25.6 percentage point lower labour force participation rate and earned $4,059 less three years after receiving benefits. Maestas et al. estimate that receiving disability insurance results in a 28 percentage point decrease in labour force participation and reduced earnings (between $3,800 and $4,600 lower) 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 themselves cause skills depreciation. Autor et al. (2015) 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. The magnitudes of the effects they find are large: the difference between rejected and accepted applicants is 27 percentage points, but this difference increases to 48 percentage points once the impact of processing time is accounted for.

Autor and Duggan (2006) investigate the effects of changes in the demand and supply of DI benefits that took place in the late 1970s and early 1980s. Supply effects arise from swings in the stringency of DI, such as the liberalisation in 1984; demand effects arise 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 programme, starting in 1984, increased disability rolls and reduced the unemployment rate among low-skill workers despite general improvements in health: the DI programme lowered measured US unemployment by 0.5 percentage points between 1984 and 2001. These effects of increased DI generosity on labour supply stem from

workers involuntarily separating from their jobs and deciding to exit the labour force, rather than from people exiting voluntarily to claim benefits.[20]

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.[21]

Elsewhere, Marie and Vall Castello (2012) exploit a discontinuity that is present in the Spanish DI programme: at age 55, Spanish DI recipients who are deemed unlikely to find work are eligible for a 36 per cent increase in their DI benefits. The authors find that labour force participation decreases substantially for those who face the benefit increase compared with 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.

b) 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 I errors) and because of acceptances onto the programme of individuals who are not disabled (type II errors). There is often a focus on the extent of type II errors but, as we now discuss, the evidence is that type I errors are the larger issue.

An early direct attempt to measure such errors in the US was made by 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 per cent of those initially awarded benefits were undeserving (type II errors), and 48 per cent of those denied were truly disabled (type I errors). Low and Pistaferri (2015) estimate these errors using a structural model

[20] Bound, Stinebrickner and Waidmann (2010) specify a dynamic programming model that looks at the interaction of health shocks, disposable income, and the labour market behaviour 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 behaviour among the old (aged 50 and over from the Health and Retirement Study).

[21] See also Anyadike-Danes and McVicar (2008).

of the DI application process and find type II errors ranging up to 18 per cent, depending on age, while type I errors are 37 per cent for those aged 45, and higher for those under 45. Similar numbers are found by Low and Pistaferri (2019) using merged administrative and survey data, with type II errors of 28 per cent and type I errors of 54 per cent. 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 Benítez-Silva, Buchinsky and Rust (2004), who use self-reported disability data on the over-50s from the Health and Retirement Study (HRS): over 20 per cent of recipients of DI are not truly work limited.

How quantitatively important these numbers are depends not just on the size of the errors but also on the numbers of healthy and disabled individuals applying. To the extent that there is already a lot of self-selection of disabled people into applying, the numbers falsely rejected are even more concerning. The incontrovertible 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 labour 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 programmes. 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 stabilised. 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. use variation in local earnings growth within states (which they argue represents 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 disability insurance programmes (measured as local spending on DI and SSI). They find sizeable elasticities of DI and SSI utilisation with respect to local earnings changes (0.35 and 0.55, respectively). Further, the authors show that participation in the DI programme 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 programmes rather than more people entering them.

Some recent papers have extended these ideas to look at the impact of the 2001 recession and of the Great Recession.[22] In particular, Maestas, Mullen and Strand (2015) show that during the Great Recession, DI applications in the US increased by 1.3 per cent for each percentage point increase in the unemployment rate; however, the award rate did not change significantly. This suggests most of the 'induced' claims came from marginal applicants.

Applications to the DI programme depend also on the implicit costs of applying. In principle, an increase in application costs may improve targeting by reducing applications from high-ability individuals with a high opportunity cost of time. On the other hand, increase in hassle may discourage applications from those most in need, in particular if behavioural considerations are important. Deshpande and Li (2019) study the effect of the closing of Social Security Administration field offices, which represents an implicit increase in the cost of applying for potential applicants residing in the states where the closings take place. The authors find that disability applications and disability awards fell by 10 per cent and 16 per cent, respectively. This discrepancy means that the closings reduce targeting efficiency, with the largest effects on those with moderately severe conditions, lower skills and lower pre-application earnings.

Evidence for other countries on the importance of benefit generosity for DI applications is similar to the US experience that is summarised in Bound and Burkhauser (1999). Mullen and Staubli (2016) use Austrian administrative data to analyse 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 receipt with respect to DI benefit generosity was 0.7, while the elasticity of applications with respect to benefits was 1.6. This implies that the screening of applications mitigates considerably the effect of increased DI generosity leading to increased numbers of DI applicants.

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 differences-in-differences 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 by an increase in return-to-work activities during sickness absence.

[22] Maestas, Mullen and Strand, 2015; Lindner, Clark and Javier, 2017.

c) 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.

In the US, the Ticket to Work programme was a policy that tried to get DI beneficiaries back to work by offering free employment services, but estimates of its effects were negligible.[23] Various states have experimented with so-called '$1 for $2' pilots, in which DI beneficiaries who work above the SGA face a reduced implicit tax rate up to some amount (a tax rate of 50 per cent instead of 100 per cent), and can keep Medicare benefits, up to some time limit. Benítez-Silva, Buchinsky and Rust (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, Song and Manchester (2011), behavioural responses to incentives in the DI programme 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 (BOND) – that implements the same '$1 for $2' format, but no results have been published at the time of writing.

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. The central conclusion from her estimates and simulations is that a significant number of individuals on DI have work capacity and are responsive to incentives to return to work, with about 30 per cent of recipients taking up work.

Evidence on the effectiveness of return-to-work incentives also exists for other countries, and estimates are obtained using a quasi-experimental setting. Kostol and Mogstad (2014) study a Norwegian return-to-work policy that was introduced in 2005. In Norway, DI beneficiaries face a 100 per cent tax rate if they work and earn above a certain threshold (the 'substantial gainful amount', or SGA). The programme studied by Kostol and Mogstad reduced the penalty for earning above the SGA; moreover, individuals awarded DI before 2004 were penalised less harshly for earning above the SGA than those whose awards came after 2004. This created a discontinuity that the authors use for identification. They find that the policy significantly increased labour force participation and earnings for recipients, did not reduce their welfare, and reduced programme costs. This suggests that many DI recipients actually have significant capacity to work.

[23] See Thornton et al. (2007).

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 programme. Exogenous variation in disincentives to stay on the programme comes partly from the age criterion and partly from the fact that the programme 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 programme. Using a regression discontinuity design, they find that the stricter reassessment rules reduced benefits on average by €1,100 per year. A small fraction of this decline (about 30 per cent) was attributable to people leaving the programme, while the bulk was due to a decline in benefits for those who remained on the programme. Next, the authors ask whether the decline in benefits was replaced by increased employment/earnings or by increased use of other social insurance programmes, 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 either by increasing employment and earnings (around two-thirds) or by more intense use of other welfare programmes (the remaining third), with little effect on the work of other family members, etc. This too supports the view that those who got cut out of the programme because of more stringent reassessment rules had substantial remaining work capacity.

2. Value of disability insurance

a) Consumption

A narrow approach to measuring the value of insurance is to consider the consumption loss associated with a negative health shock. This is done by Gallipoli and Turner (2009) and Meyer and Mok (2019) in the US and by Ball and Low (2014) in the UK. Consumption in both the UK and the US falls on disability, but these declines are mitigated by the receipt of disability insurance. However, as discussed by Ball and Low, there is selection into who actually receives DI 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. Ball and Low estimate consumption losses of 9 per cent for those in receipt of DI, but lower falls for those not in receipt, reflecting this selection effect. This implies that estimates are lower bounds of the true mitigation achieved by DI. Meyer and Mok use the estimates of consumption loss in a Baily–Chetty framework[24] to consider the optimality of making DI more generous. What this approach

[24] Baily, 1978; Chetty, 2008.

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. Instead, this approach evaluates the insurance value gain of an increase in generosity at the margin.

Deshpande, Gross and Su (2019) look at the impact of outcomes of disability insurance applications on financial distress, such as the likelihood of declaring personal bankruptcy, foreclosure or being evicted. The award of disability insurance substantially mitigates these indicators of distress, and ignoring these effects underestimates the value of disability insurance.

Michaud and Wiczer (2018) argue that, in principle, disability insurance is valuable not only for the usual risk sharing reasons (consumption smoothing), but also because of more efficient occupational reallocation – more workers would choose to work in risky occupations (hence increasing output) relative to a case where no disability insurance existed. Nevertheless, in their calibrated model (using differences in disability risks across occupations), the gains in welfare from moving to an optimal DI generosity are mostly from risk sharing, not production efficiency.

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 programmes that mitigate the risk, accounting for the fiscal cost of the programme and for the changes in labour supply and saving behaviour resulting from the programme. These calculations clearly involve a cardinalisation of preferences and introducing substantial structure.

Work by Waidmann, Bound and Nichols (2003), Bound et al. (2004), Benítez-Silva, Buchinsky and Rust (2005), Bound, Stinebrickner and Waidmann (2010) and Low and Pistaferri (2015) has highlighted the importance of considering both sides of the insurance–incentive trade-off for welfare analysis and has conducted some policy experiments evaluating the consequences of reforming the DI programme. These papers differ in focus, which leads to differences in the way preferences, risk and the screening process are modelled, and in the data and estimation procedure used.[25] In Section IV, we summarise our own work on the subject[26] to provide an explicit framework for thinking about these trade-offs and the choices of structure that need to be made.

[25]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.

[26]Low and Pistaferri, 2015.

b) Spousal labour supply, savings and other insurance

The value of disability insurance may depend on the availability of other sources of insurance, such as other welfare programmes or the labour supply of other household members, through added-worker effects.

Mueller, Rothstein and von Wachter (2016) study the interactions between the UI and the DI programmes after the Great Recession, showing that DI applications are countercyclical. During the Great Recession, the UI programme was extended considerably from the statutory 26 weeks to up to 99 weeks. In principle, this would have slowed down applications to the DI programme if the two programmes 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. Schmidt, Shore-Sheppard and Watson (2019) reach a similar conclusion examining the expansion of Medicaid for those not in receipt of SSI or DI. This expansion reduces the incentive to go onto SSI/DI, but the authors find no evidence of any effect. On the other hand, Goodman-Bacon and Schmidt (2019) show that the introduction at federal level of SSI in 1974 shrank non-disability cash transfer programmes, such that each extra $1 of spending on disability increased total spending by only 50 cents.

Deshpande (2016) investigates the effect of cessation in SSI payments to children on their parents' labour 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 programme. She finds that parents respond to being dropped from the programme by increasing labour effort, approximately offsetting the children's lost disability income one-for-one. Furthermore, they do not generally substitute towards other transfer programmes.

Autor et al. (2017) study the added-worker effects associated with disability insurance using Norwegian administrative data and a 'judge assignment' research design. They find that DI increases household income and consumption significantly for single-person households. They also find that DI receipt reduces usage of other transfer programmes, but by less than one-for-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 his or her labour decisions in response to DI decisions.[27] This is in contrast with Gallipoli and Turner (2009), who find little evidence of added-worker effects in response to disability onset.

[27] Lee (2019) explores substitution towards informal care by spouses and finds evidence that this attenuates added-worker insurance effects (raising the total value of DI). Fadlon and Nielsen (2017) use administrative Danish data and show that non-fatal health shocks such as heart attacks or strokes suffered by one spouse have no meaningful effects on spousal labour supply, consistent with the adequate insurance coverage for the associated forgone income.

IV. Organising framework: insurance and incentives

The empirical literature discussed in the previous section has highlighted the importance of disincentives to work and the welfare value of disability insurance programmes. To fully understand the interplay between the insurance and incentive consequences and to evaluate the welfare effect of reforms, a theoretical framework is needed. The difficulty with modelling disability insurance is that the details of the programmes vary substantially, as seen in Section II. Further, by contrast with UI, consideration of health and health shocks can impact on individuals in multiple ways. In thinking about UI, researchers now often use the Baily–Chetty framework, discussed by Spinnewijn (2020; this issue). Meyer and Mok (2019) apply this same framework to disability insurance, but, as mentioned above, 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 individuals' resources 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 discuss what we believe should be the key ingredients of an organising theoretical framework. Additionally, we discuss the substantial measurement issues to consider and we re-evaluate the empirical literature in its light. In Section V, we will use this framework for understanding policy reform.

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 and some false applicants are accepted. 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 do 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?

a) Modelling preferences

We leave to one side the issue of how decisions within a family are made, and consider an individual maximising 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?', which one can write as

$$(1) \qquad U_t = E_t \sum_{s=t}^{T} \beta^{s-t} u(c_t, l_t, d_t),$$

where β is the discount factor, E_t is 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 equalised 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 Panel Study of Income Dynamics (PSID) consumption data and self-reported disability data. On the other hand, Finkelstein, Luttmer and Notowidigdo (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 holidays), while others are complements (such as alternative transportation services and domestic services). Further, it is not clear that some forms of expenditure give utility directly, and instead serve only to change health through some production process.

b) Modelling constraints and resources

We discuss modelling the impact of health on private resources in this subsection and modelling the mitigating impact of government insurance in the

next. 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 job destruction rates.

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

$$\ln w_{it} = x'_{it}\beta + \alpha d_{it} + \zeta_{it}, \tag{2}$$

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 or the price of their skills have deteriorated permanently due to automation, international trade effects, etc., this makes the opportunity cost of applying for disability insurance low. Further, 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 labour market frictions: people with marginal disabilities who face higher job destruction rates or lower arrival rates of employment offers are more likely to become applicants to the programme. In the empirical literature, several papers have studied such effects.[28] 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 modelled 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[29] 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, not smoking, low-fat diets, etc. more than low-human-capital individuals. A confounder is that high education is also correlated with working in occupations and industries where the likelihood of developing a work-related disability is lower. In these more complex models where disability evolves endogenously, consumption choices, such as of out-of-pocket healthcare spending, could

[28] For example, Black, Daniel and Sanders (2002).
[29] As in Grossman (1972).

be seen as an input in the production function of health investments which augment (or replace depreciated) health capital.

c) Modelling government insurance

One way to insure against disability risk is self-insurance through precautionary saving. However, since disability is a low-probability event with (usually) catastrophic consequences, self-insurance against disability risk is rarely if ever efficient. As discussed in Autor et al. (2017), it is 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.[30] The challenge is how to structure the complex links between the various social insurance programmes people have access to in the US, the UK and other countries. There are two main challenges in thinking about models of disability insurance: first, the programmes have complex rules, as documented in Section II; and second, there are important interactions between different social insurance programmes.

If application to the disability insurance programme is modelled as an explicit choice, the elements described in Section II 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 reassessment process. All of these change the decision to apply for benefits or to stay on the programme. One option is to model directly the 'supply side', i.e. take a stand on how the SSA models the screening process (the extent of signal/noise it observes, the disability threshold it sets, the resources available for reassessment, etc.). Another option is to take a more reduced form approach in which the probability of an award is a function of the applicant's 'type', defined by skills, age and the extent of actual work limitations, plus a stochastic component. Skills and age enter directly in the screening process (especially at the 'vocational' stage), while the extent of work limitations captures 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 the 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 benefits. Social insurance is through unemployment insurance, which is a temporary programme 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. The welfare programmes included are a basic means-tested

[30] About 50 per cent 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 (2014)).

income-support programme (such as food stamps) and a means-tested programme such as SSI that combines the presence of disability with low income. The division of these various programmes into 'social insurance' programmes and 'welfare' programmes is because of differences in contribution requirements before claiming, but in practice all four programmes provide insurance against different sorts of shocks: welfare programmes 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 criterion for DI is examined, since it too contains a requirement about not having appropriate skills for work, as discussed in Section II.

The consequences of these programmes are intertwined. For example, a more generous UI could disincentivise applications to the DI programme during economic busts, as may have happened during the early phases of the Great Recession in the US.[31] 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 labour market. Modelling the links between these programmes correctly is important for judging the success of reforms to disability insurance programmes, as such reforms typically impact take-up of alternative programmes.

d) 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 attention to the detail of the programme and the choices of individuals. In Section III, we reported evidence on the two approaches.

2. Measurement issues

As argued above, a key issue in modelling how health shocks and disability insurance impact behaviour is how to measure the extent of work disabilities. This is the heart of the 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 programmes, need measures of disability that are as close as possible to the one adopted by the SSA, as quoted in Section II. In principle, disability d is a continuous measure of disability (and,

[31] Mueller, Rothstein and von Wachter, 2016.

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, HRS and the Survey of Income and Program Participation, or SIPP), 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 indicator to classify people as work disabled. Some data sets (such as PSID) ask follow-up questions to those answering 'Yes', such as 'Does this condition keep you from doing some types of work?' (possible answers being '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').[32] 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 enables 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 with the simple binary question.[33]

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 2019) provide additional evidence using PSID and HRS data, respectively.

Second, individuals may overestimate their work limitation in order to justify their disability payments or their non-participation in the labour force. Benítez-Silva, Buchinsky and Rust (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.

[32]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 answers 'Yes' to the first and second questions and 'Not temporary' to the third question.

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

Third, health status may be endogenous, and non-participation in the labour force may affect health (either positively or negatively). Stern (1989) and Bound (1989) both find positive effects of non-participation on health, but the effects are economically small. Further, Smith (2004) finds that income does not affect health once one controls for education (as Low and Pistaferri (2015) do implicitly by focusing on a group of homogeneous individuals with similar schooling levels). Similarly, Adda, Banks and von Gaudecker (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 interpersonal comparability. Some researchers have pioneered the use of disability vignettes[34] 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 interpersonal differences in subjective disability thresholds.[35]

V. Policy implications

Given the accumulating evidence on the distortions induced by increasing the generosity of disability insurance, as well as the undeniable importance of insurance for those who are genuinely hit by disability shocks, we conclude this survey with a discussion of the policy implications of potential reforms. We consider the evidence on the policy dimensions outlined in Section II.

We report results mostly from our previous work, Low and Pistaferri (2015). In that work, we stress that a life-cycle perspective offers a useful way to capture fully the insurance benefits. We also argue that capturing fully the incentive costs of the programme requires an accurate characterisation both of labour supply behaviour and of applications to the programme. 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

[34] See Kapteyn, Smith and van Soest (2007).

[35] 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 has been done in the education literature (see Cunha, Heckman and Schennach (2010)).

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 programme 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 programme 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. To calculate the welfare implications of the various reforms described below, Low and Pistaferri (2015) measure the willingness to pay (in consumption terms) for the new policy, i.e. the fraction of consumption that would make an 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 we abstract from general equilibrium effects.

1. Medical tests and stringency

A first reform to consider is changing the strictness of the screening process. In the US, a reform of this type 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 centres. Low and Pistaferri (2015) find that increasing the stringency of the screening process by raising the disability threshold for admission into the programme reduces the extent of the incentive problem, but also reduces the extent of insurance provided by the programme as expected. This reduces welfare (expected utility) and is a clear example of the trade-off between incentives and insurance.[36] 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 of getting into the programme is highly valued.

2. Eligibility requirements

As stressed above, disability insurance programmes may interact in important ways with other government welfare programmes, such as food stamps or

[36] An alternative policy might be to reduce the noise involved in the evaluation of the signal, but this policy may be either too costly or unfeasible.

income support. Low and Pistaferri (2015) investigate the importance of such interactions by changing the generosity of the means-tested programme. Marginal applicants to DI switch their decision from applying to not-applying: for 'false applicants', the means-tested programme acts as a *substitute* for DI and generally applications to DI fall as income support generosity increases. 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 becomes 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 income support 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 labour force and to stop working altogether.

3. Generosity and progressivity

One of the lessons from the trade-offs captured in Low and Pistaferri (2015) is that increases in DI generosity can be welfare improving despite the increase in moral hazard (false applications) it generates.[37] 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 a higher tax rate on workers), although the effect varies substantially with the underlying productivity of individuals.

4. 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 programme more than the cash benefits component of the programme. In the era before the Affordable Care Act (ACA), 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 seven

[37]Meyer and Mok (2019) reach a similar conclusion. They apply a variant of the benefit optimality formula derived by Baily (1978) and 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.

years following their nine-month trial work period. However, take-up rates for this programme have been substantially low. There is no research to date on the impact of ACA (which eliminated denials of private health insurance because of pre-existing conditions) on DI application rates. On the other hand, we have reported evidence that those on disability have residual work capabilities.

5. Concluding thoughts

In terms of the accuracy of the DI programmes, the main lesson is the substantial false rejections of those who are in need of insurance. Only 58 per cent of the severely work limited in the US 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 programme, coupled with labour market rehabilitation from the moment of application, is a more effective way of providing insurance for those in need alongside minimising the extent of false recipients. There are no simple solutions to the insurance–incentive trade-off. However, creative programme design can reduce incentive costs and improve insurance, as shown by the Netherlands through the 2000s.

References

Acemoglu, D. and Angrist, J. D. (2001), 'Consequences of employment protection? The case of the Americans with Disabilities Act', *Journal of Political Economy*, vol. 109, pp. 915–57.

Adda, J., Banks, J. and von Gaudecker, H-M. (2009), 'The impact of income shocks on health: evidence from cohort data', *Journal of the European Economic Association*, vol. 7, pp. 1361–99.

Anyadike-Danes, M. and McVicar, D. (2008), 'Has the boom in incapacity benefit claimant numbers passed its peak?', *Fiscal Studies*, vol. 29, pp. 415–34.

Autor, D. and Duggan, M. (2006), 'The growth in the social security disability rolls: a fiscal crisis unfolding', *Journal of Economic Perspectives*, vol. 20, no. 3, pp. 71–96.

—, — and Gruber, J. (2014), 'Moral hazard and claims deterrence in private disability insurance', *American Economic Journal: Applied Economics*, vol. 6, no. 4, pp. 110–41.

—, Kostol, A. R., Mogstad, M. and Setzler, B. (2017), 'Disability benefits, consumption insurance, and household labor supply', National Bureau of Economic Research (NBER), Working Paper no. 23466.

—, Maestas, N., Mullen, K. and Strand, A. (2015), 'Does delay cause decay? The effect of administrative decision time on the labor force participation and earnings of disability applicants', National Bureau of Economic Research (NBER), Working Paper no. 20840.

Baily, M. N. (1978), 'Some aspects of optimal unemployment insurance', *Journal of Public Economics*, vol. 10, pp. 379–402.

Ball, S. and Low, H. (2014), 'Do self-insurance and disability insurance prevent consumption loss on disability?', *Economica*, vol. 81, pp. 468–90.
Banks, J., Blundell, R. and Emmerson, C. (2015), 'Disability benefit receipt and reform: reconciling trends in the United Kingdom', *Journal of Economic Perspectives*, vol. 29, no. 2, pp. 173–90.
Bell, B. and Smith, J. (2004), 'Health, disability insurance and labour force participation', Bank of England, Working Paper no. 218.
Benítez-Silva, H., Buchinsky, M., Chan, H. M., Rust, J. and Sheidvasser, S. (1999), 'An empirical analysis of the social security disability application, appeal, and award process', *Labour Economics*, vol. 6, pp. 147–78.
—, — and Rust, J. (2004), 'How large are the classification errors in the social security disability award process?', National Bureau of Economic Research (NBER), Working Paper no. 10219.
—, — and — (2005), 'Induced entry effects of a $1 for $2 offset in SSDI benefits', Stony Brook University, Department of Economics, Working Paper no. 05-03.
—, Disney, R. and Jiménez-Martín, S. (2010), 'Disability, capacity for work and the business cycle: an international perspective', *Economic Policy*, vol. 25, pp. 483–536.
Bianco, C. D. (2019), 'Disability insurance and the effects of return-to-work policies', University of Padua, mimeo.
Black, D., Daniel, K. and Sanders, S. (2002), 'The impact of economic conditions on participation in disability programs: evidence from the coal boom and bust', *American Economic Review*, vol. 92, pp. 27–50.
Borghans, L., Gielen, A. C. and Luttmer, E. F. P. (2014), 'Social support substitution and the earnings rebound: evidence from a regression discontinuity in disability insurance reform', *American Economic Journal: Economic Policy*, vol. 6, no. 4, pp. 34–70.
Bound, J. (1989), 'The health and earnings of rejected disability insurance applicants', *American Economic Review*, vol. 79, pp. 482–503.
— and Burkhauser, R. V. (1999), 'Economic analysis of transfer programs targeted on people with disabilities', in O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, vol. 3, Amsterdam: Elsevier.
—, Cullen, J. B., Nichols, A. and Schmidt, L. (2004), 'The welfare implications of increasing disability insurance benefit generosity', *Journal of Public Economics*, vol. 88, pp. 2487–514.
—, Stinebrickner, T. and Waidmann, T. (2010), 'Health, economic resources and the work decisions of older men', *Journal of Econometrics*, vol. 156, pp. 106–29.
Burkhauser, R. and Daly, M. (1996), 'Employment and economic well-being following the onset of a disability: the role for public policy', in J. Mashaw, V. Reno, R. Burkhauser and M. Berkowitz (eds), *Disability, Work and Cash Benefits*, Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.
— and — (2011), *The Declining Work and Welfare of Working-Age People with Disabilities*, Washington DC: AEI Press.
Case, A. and Deaton, A. (2017), 'Mortality and morbidity in the 21st century', *Brookings Papers on Economic Activity*, Spring, pp. 397–443.
Chen, S. and van der Klaauw, W. (2008), 'The work disincentive effects of the disability insurance program in the 1990s', *Journal of Econometrics*, vol. 142, pp. 757–84.
Chetty, R. (2008), 'Moral hazard versus liquidity and optimal unemployment insurance', *Journal of Political Economy*, vol. 116, pp. 173–234.
Cunha, F., Heckman, J. J. and Schennach, S. M. (2010), 'Estimating the technology of cognitive and noncognitive skill formation', *Econometrica*, vol. 78, pp. 883–931.
de Jong, P., Lindeboom, M. and van der Klaauw, B. (2011), 'Screening disability insurance applications', *Journal of the European Economic Association*, vol. 9, pp. 106–29.

DeLeire, T. (2000), 'The wage and employment effects of the Americans with Disabilities Act', *Journal of Human Resources*, vol. 35, pp. 693–715.
Deshpande, M. (2016), 'The effect of disability payments on household earnings and income: evidence from the SSI children's program', *Review of Economics and Statistics*, vol. 98, pp. 638–54.
—, Gross, T. and Su, Y. (2019), 'Disability and distress: the effect of disability programs on financial outcomes', National Bureau of Economic Research (NBER), Working Paper no. 25642.
— and Li, Y. (2019), 'Who is screened out? Application costs and the targeting of disability programs', *American Economic Journal: Economic Policy*, vol. 11, pp. 213–48.
Diamond, P. and Sheshinski, E. (1995), 'Economic aspects of optimal disability benefits', *Journal of Public Economics*, vol. 57, pp. 1–23.
Fadlon, I. and Nielsen, T. H. (2017), 'Family labor supply responses to severe health shocks', National Bureau of Economic Research (NBER), Working Paper no. 21352.
Finkelstein, A., Luttmer, E. F. P. and Notowidigdo, M. J. (2013), 'What good is wealth without health? The effect of health on the marginal utility of consumption', *Journal of the European Economic Association*, vol. 11, suppl. 1, pp. 221–58.
French, E. and Song, J. (2014), 'The effect of disability insurance receipt on labor supply', *American Economic Journal: Economic Policy*, vol. 6, pp. 291–337.
Gallipoli, G. and Turner, L. (2009), 'Household responses to individual shocks: disability and labour supply', mimeo.
Golosov, M. and Tsyvinski, A. (2006), 'Designing optimal disability insurance: a case for asset testing', *Journal of Political Economy*, vol. 114, pp. 257–79.
Goodman-Bacon, A. and Schmidt, L. (2019), 'Federalizing benefits: the introduction of supplemental security income and the size of the safety net', National Bureau of Economic Research (NBER), Working Paper no. 25962.
Grossman, M. (1972), *The Demand for Health: A Theoretical and Empirical Investigation*, New York, NY: Columbia University Press.
Gruber, J. and Kubik, J. D. (1997), 'Disability insurance rejection rates and the labor supply of older workers', *Journal of Public Economics*, vol. 64, pp. 1–23.
Kapteyn, A., Smith, J. P. and van Soest, A. (2007), 'Vignettes and self-reports of work disability in the United States and the Netherlands', *American Economic Review*, vol. 97, pp. 461–73.
Kostol, A. R. and Mogstad, M. (2014), 'How financial incentives induce disability insurance recipients to return to work', *American Economic Review*, vol. 104, pp. 624–55.
Kreider, B. (1999), 'Latent work disability and reporting bias', *Journal of Human Resources*, vol. 34, pp. 734–69.
Lakdawalla, D., Bhattacharya, J. and Goldman, D. (2004), 'Are the young becoming more disabled?', *Health Affairs*, vol. 23, pp. 168–76.
Lee, S. (2019), 'Household responses to disability shocks: spousal labor supply, caregiving, and disability insurance', mimeo.
Lillard, L. and Weiss, Y. (1997), 'Uncertain health and survival: effects on end-of-life consumption', *Journal of Business and Economic Statistics*, vol. 15, pp. 254–68.
Lindner, S., Clark, B. and Javier, M. (2017), 'Characteristics and employment of applicants for social security disability insurance over the business cycle', *B. E. Journal of Economic Analysis and Policy*, vol. 17, no. 1.
Low, H. and Pistaferri, L. (2015), 'Disability insurance and the dynamics of the incentive insurance trade-off', *American Economic Review*, vol. 105, pp. 2986–3029.
— and — (2019), 'Disability insurance and gender differences: evidence from merged survey-administrative data', University of Oxford, mimeo.

Maestas, N., Mullen, K. J. and Strand, A. (2013), 'Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt', *American Economic Review*, vol. 103, pp. 1797–829.
—, — and — (2015), 'Disability insurance and the Great Recession', *American Economic Review*, vol. 105, pp. 177–82.
Marie, O. and Vall Castello, J. (2012), 'Measuring the (income) effect of disability insurance generosity on labour market participation', *Journal of Public Economics*, vol. 96, pp. 198–210.
Meyer, B. D. and Mok, W. K. (2019), 'Disability, earnings, income and consumption', *Journal of Public Economics*, vol. 171, pp. 51–69.
Michaud, A. and Wiczer, D. (2018), 'Occupational hazards and social disability insurance', *Journal of Monetary Economics*, vol. 96, pp. 77–92.
Morden, N., Munson, J., Colla, C., Skinner, J., Bynum, J., Zhou, W. and Meara, E. (2014), 'Prescription opioid use among disabled Medicare beneficiaries: intensity, trends, and regional variation', *Medical Care*, vol. 52, pp. 852–9.
Mueller, A. I., Rothstein, J. and von Wachter, T. M. (2016), 'Unemployment insurance and disability insurance in the Great Recession', *Journal of Labor Economics*, vol. 34, pp. S445–75.
Mullen, K. J. and Staubli, S. (2016), 'Disability benefit generosity and labor force withdrawal', *Journal of Public Economics*, vol. 143, pp. 49–63.
Nagi, S. (1969), *Disability and Rehabilitation*, Columbus, OH: Ohio State University Press.
Parsons, D. O. (1980), 'The decline in male labor force participation', *Journal of Political Economy*, vol. 88, pp. 117–34.
Rupp, K. and Stapleton, D. C. (eds) (1998), *Growth in Disability Benefits: Explanations and Policy Implications*, Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.
Schmidt, L., Shore-Sheppard, L. and Watson, T. (2019), 'The impact of the ACA Medicaid expansion on disability program applications', National Bureau of Economic Research (NBER), Working Paper no. 26192.
Smith, J. P. (2004), 'Unraveling the SES–health connection', *Population and Development Review*, vol. 30, pp. 108–32.
Spinnewijn, J. (2020), 'The trade-off between insurance and incentives in differentiated unemployment policies', *Fiscal Studies*, vol. 41, pp. 101–127.
Stern, S. (1989), 'Measuring the effect of disability on labor force participation', *Journal of Human Resources*, vol. 24, pp. 361–95.
Thornton, C., Livermore, G., Fraker, T., Stapleton, D., O'Day, B., Witternburg, D., Weathers, R., Goodman, N., Silva, T., Sama Martin, E., Gregory, J., Wright, D. and Mamun, A. (2007), *Evaluation of the Ticket to Work Program: Assessment of Post-Rollout Implementation and Early Impacts*, Washington DC: Mathematica Policy Research.
von Wachter, T., Song, J. and Manchester, J. (2011), 'Trends in employment and earnings of allowed and rejected applicants to the social security disability insurance program', *American Economic Review*, vol. 101, pp. 3308–29.
Waidmann, T., Bound, J. and Nichols, A. (2003), 'Disability benefits as social insurance: tradeoffs between screening stringency and benefit generosity in optimal program design', Michigan Retirement Research Center, Working Paper no. 2003-042.
Wu, A. Y. and Hyde, J. S. (2019), 'The postretirement well-being of workers with disabilities', *Journal of Disability Policy Studies*, vol. 30, pp. 46–55.