Disability Insurance: Error Rates and Gender Differences *

Hamish Low† and Luigi Pistaferri‡

July 30, 2020

Abstract

We show the extent of errors made in the award of disability insurance using matched survey-administrative data. False rejections (Type I errors) are widespread and characterized by large gender differences. Women with a severe, work-limiting, permanent impairment are 20 percentage points more likely to be rejected than men, controlling for the type of health condition, occupation, and a host of demographic characteristics. The differences by gender arise because women are more likely to be assessed as being able to find other work than observationally equivalent men. Despite this, after rejection, women with a self-reported work limitation do not return to work, compared to rejected women without a work limitation. We investigate whether these gender differences in Type I errors are due to women being in better health than men, to women having lower pain thresholds, to women applying more readily for disability insurance, or to women applying with harder-to-verify work limitations. None of these explanations are consistent with the data. By contrast, we find evidence suggesting that there are different acceptance thresholds for men and women.

Keywords: Disability Insurance, Gender Differences.
JEL Classification Codes: I38, J16.

*Thanks to Sarah Eichmeyer and Maxwell Rong for invaluable research assistance, and to David Card, David Autor, Stephen Haider, Hanming Fan, Amanda Michaud, Magne Mogstad, John Rust, Lucie Schmidt, Alessandra Voena, and seminar participants at Berkeley, Bonn, Chicago, Georgetown, Michigan State, Minnesota, Oxford, Pavia, Penn, Royal Holloway, Sheffield, Surrey, the IFS Conference on "Earnings, Risk and Insurance", the NBER Labor Winter meeting, and the C6 Conference in Capri for comments. This paper uses restricted HRS data made available to Pistaferri under a confidential agreement. All errors are ours.

†University of Oxford and Institute for Fiscal Studies.
‡Stanford University, SIEPR, NBER and CEPR.

1 Introduction

Disability insurance is now a major part of social insurance provision in the US and elsewhere. There is substantial evidence on the labor supply incentive effects of the program. There is however very little evidence on how well targeted the program is and what errors are being made through the award process. Further, there is no evidence at all on whether these errors differ by gender or other observable characteristics. The aim of this paper is to fill this gap, to show the extent of false rejections and false acceptances, and how these differ between men and women.

We focus on the two major programs in the US that pay benefits against disability risk for working age individuals: the Social Security Disability Insurance (DI) program and the Supplemental Security Income (SSI) program. Both programs are evaluated by the Social Security Administration (SSA) using the same medical assessment process to determine disability and eligibility.[1] The size of these programs has generated concerns that some working-able individuals may exaggerate their disability in order to become beneficiaries or quit into unemployment in order to apply for benefits; and further, that beneficiaries may have little incentive to go back to work even when their health condition improves.[2] Reflecting these concerns, there exists a sizable literature that has looked at the labor supply incentive consequences of disability insurance.[3] These concerns about labor supply consequences and false applications arise because the true disability status of an applicant is unknown. This in turn means that SSA examiners are prone to making two type of errors: Type I errors (rejecting a truly disabled applicant) and Type II errors (awarding benefits to applicants who are not truly disabled).

The extent of these inefficiencies may depend on the applicants' characteristics, such

[1]Both programs have attracted a lot of attention in recent years (see Duggan and Imberman, 2009, for a survey) due to their cost and the fast increases in case-loads. The difference between the programs is that the DI program is financed through payroll taxation and pays benefits to covered workers, whereas the SSI program is financed through general taxation and pays benefits to low-income individuals. In 2017 the DI program was paying cash benefits of around $134 billion (in comparison, the Unemployment Insurance (UI) program was paying benefits worth only $28 billions). In the same year, total SSI expenditure was $59 billion, absorbing 16% of federal non-Medicaid welfare spending. In terms of growth of recipiency, between 1984 and 2017 the share of disabled workers receiving DI benefits out of all workers increased from 2.4% to 5.6%. As for SSI, the fraction of 18-64 years old who are receiving SSI benefits has doubled from 1.2% in 1984 to 2.4% in 2017.

[2]In the case of SSI, the additional concern is that applicants may be discouraged from saving in order to meet the asset test.

[3]Some of the earlier empirical literature is surveyed in Bound and Burkhauser (1999) and Haveman and Wolfe (2000). More recent contributions are surveyed in Low and Pistaferri (2019).

as the particular health condition. Classification errors may be higher for conditions that are harder to verify, such as muscolosketal or mental disorders. Indeed, the SSA disability determination process distinguishes explicitly between individuals with a so-called "listed impairment" which gives automatic qualification, and those without where both the nature of the health condition and work adaptability are taken into account. Further, besides health, the SSA process explicitly makes the probability of award a function of age, occupation and work experience, as discussed by Chen and van der Klaauw (2008).[4]

At the heart of the problem the SSA faces is that disability assessment is subjective and this opens up the possibility of bias, even if unintended. Alternatively, as discussed in the medical literature (Legato et al., 2016; Bangasser et al., 2019; Clocchiatti et al., 2016), gender differences may arise because women with a given disability, present symptoms in a different way from men and this may affect disability insurance screening.[5] There are now several examples in other contexts in which subjective assessment leads to bias on the basis of gender. Card et al. (2019) show that different standards are imposed by the editorial process on female authors in economics journals. Sarsons (2019) shows that the performance of a surgeon is evaluated differently by the referring doctor depending on the gender of the surgeon. The context of disability insurance is a case where the consequences of bias are potentially severe: those rejected for disability insurance despite not being able to work, have often very few alternative avenues of support, and this is a long-term problem.

The difficulty in estimating Type I and Type II errors is that we need measures of "true" health-related work limitations alongside "reported to the authorities" work limitations. We solve this difficulty by merging information on self-reported disability from the Health and Retirement Study (HRS) with administrative data from the SSA on DI and SSI applications and social security earnings.[6] This enables us to estimate directly the extent of Type I and Type II errors because we have independent data on the "true" work-limitation as well as

[4]Given the long waiting periods and different appeal processes that applicants go through to get onto DI or SSI, we might expect different stages of application to be subject to different sorts of error. Further, there is an interaction between effects: work limitations caused by muscoloskeletal or mental health conditions tend only to lead to disability awards at the later stages of the appeal process.

[5]It is also possible that the screening system evolves (with lags) to fit the gender composition of applicants, who were initially mostly men. However, this is rapidly changing, with women representing in 2016 almost half of the stock and half of the flow of new entrants into DI (up from 1/3 and 1/4 in the mid 1980's).

[6]While DI and SSI have mostly been studied in isolation, it may be valuable to study them jointly because the formal definition of disability is the same in both programs and the disability determination process is done by the same agencies and officers (local Social Security field offices). Merging application data from the two programs makes inference more reliable because in survey data the number of applicants to either program is typically small.

the report of the SSA. To the best of our knowledge, this is the first paper that attempts to use these linked data for studying the efficiency aspects of the DI/SSI programs.

Armed with these measures, we study whether these errors differ by observable characteristics of applicants. We document significant gender differences in Type I error rates. Women with a severe, work-related, permanent impairment are more likely to have their disability insurance application turned down (i.e., suffer a Type I error) than men with observationally equivalent characteristics. This main finding is robust to numerous sensitivity checks. The point at which this difference by gender arises is not in the assessment of whether a health condition exists, but rather in the assessment of whether any given medical condition prevents the applicant from doing alternative work. Among men and women with the same health condition, women are more likely to be assessed as being able to find other work. This difference arises across health conditions of differing severity and verifiability. The conclusion that the SSA believes the medical condition of disabled women does not inhibit their labor market choices can be tested directly by looking at labor market outcomes after rejection. We find that among rejected women, those who previously self-reported severe work-limitations return to work at a much lower rate than those that do not report a work-limitation. This suggests that the SSA has wrongly assessed these women as being able to work. Indeed, no such differences emerge among men or among women in the years preceding application.

We propose a simple model of why Type I errors may differ by gender. Possible explanations are that men and women could differ in terms of severity of actual or perceived work-related impairments, or differ in terms of opportunity costs of applying. On the supply side of benefits, men and women could face different admission standards, for example because evaluators use a "unisex" approach that disadvantages women, or because the noise of health signals differ by gender. To assess these explanations, we use data on self-reports of work limitations, application rates and rejection rates. We also use survey respondents' assessments of the work limitations of individuals described in disability vignettes. The gender of the individuals described in the vignettes is randomized and this provides insights on how disabled men and women are assessed differently. Our conclusion is that supply considerations, and in particular differences in the award thresholds for men and women, are more plausible explanations than demand-side channels for the gender differences.

There are only a very few papers that estimate classification errors associated with disability insurance. One early study is Nagi (1969), where a sample of 2,454 DI applicants were assessed by a team of medical professionals independent of the SSA and this assessment

was compared to the SSA assessment. Nagi (1969) concluded that, at the time of the award, about 19% of those initially awarded benefits were undeserving, and 48% of those denied were truly disabled. The obvious limitation of the Nagi (1969) study is that this refers to a period in which the disability programs were fundamentally different (indeed, the SSI started only in 1974). The most dramatic difference since then was the 1984 Social Security Disability Benefits Reform Act that liberalized admission criteria for DI and SSI, resulting in a large increase in applicants and people awarded benefits with mental health and musculoskeletal conditions.[7] Since these are hard-to-verify conditions, classification errors in the post-1984 era may be very different. This also highlights that errors may differ by health condition.

To the extent that individuals recover but do not flow off DI, we would expect the fraction falsely claiming to be higher in the stock than at admission. This is the finding of Benitez-Silva et al. (2004) who use the self-reported binary indicator of work limitations in the HRS as a classical error-ridden measure of the "true" disability status, and compare this to the reported outcome of a self-reported DI/SSI application. They compute classification errors for DI and SSI combined and find that over 40% of recipients of DI/SSI are not truly work limited.

Low and Pistaferri (2015) follow a similar strategy of using self-reported work limitations alongside details of receipt of DI, taking data from the Panel Study of Income Dynamics (PSID). They distinguish between severe and moderate work disability instead of using a binary indicator, and estimate classification errors using a structural model to capture the application decision. Similarly to Nagi (1969), Low and Pistaferri (2015) find that the Type I error is large (approximately 2/3 for younger workers and 1/3 for older workers), while the Type II error is concentrated among those with moderate disabilities (18%) with the error being only 1% among those who apply while reporting no disabilities.

There are several issues that make estimates of classification errors from the studies above problematic. First, how strong is the "signal" embedded in the self-reported disability measures and how well does it correspond to the SSA assessment criteria. Second, survey data relies on recall data of the application process, which may be subject to measurement

[7] Among the provisions of the Act there were at least three that may have increased the probability of admission because of such conditions: (1) the requirement that SSA obtain evidence from the applicant's treating physician instead of hired consultants, "since the treating physician is likely to be the medical professional most able to provide a detailed, longitudinal picture of the individual's medical condition"; (2) updating the criteria for evaluating mental impairments to "make them consistent with present-day diagnosis, treatment, and evaluation"; (3) the requirement that the SSA, in determining the severity of a person's disability, "consider the combined effect of all impairments without regard to whether any one impairment, if considered separately, would be severe" (Collins and Herfle, 1985).

(recall) errors. These are particularly relevant in cases in which a disability improves or worsens, since one needs to "pin" the disability status at the time of the application in order to assess the extent of classification errors. Moreover, in some years of the HRS, disability application questions are only asked to those who report having a disability, which induces a mechanical understatement of Type II errors. The key limitation of these papers however is that none try to understand how classification errors vary by how verifiable the health condition is and whether errors vary by demographic characteristics.[8]

The rest of the paper proceeds as follows. In Section 2 we provide institutional details on the programs that insure against work limitation shocks. Section 3 presents a simple theoretical framework to guide through the empirical findings on error rates. After presenting the data in section 4, we discuss the results in Section 5. Section 6 tests the alternative mechanisms that may be behind our findings. Section 7 concludes.

2 Institutional Details

2.1 The DI program

The DI program is a social insurance program that provides cash and health care benefits for covered workers, their spouses, and dependents. The purpose of the program is to provide insurance against persistent health shocks that impair substantially the ability to work. In other words, the assessment is a combination of health and residual work capacity.[9] The difficulty with providing insurance is, of course, that health status and the impact of health on the ability to work are imperfectly observed. Cash benefits are computed using the same formulae used to compute Social Security retirement benefits.[10] While benefits are independent of the extent of the work limitation, caps on the payroll tax financing the DI program as well as the nature of the formula determining benefits make the system progressive. Because of the progressivity of the benefits and because individuals receiving

[8] A study by the United States General Accounting Office (1994) reported higher disability insurance denial rates among women. Their explanation was that a significant fraction of the difference could be explained by occupation dummies and the fact that SSA evaluators assessed that women apply with lower impairments than men. However, the study had no measures of actual health conditions independent of the SSA.

[9] The emphasis on the severity and persistence of the health shock distinguishes the DI program from the Workers Compensation program, which pays cash and health care benefits for temporary health shocks that are work-related, or private medical leave programs.

[10] 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).

DI also receive Medicare benefits after two years, the replacement rates are substantially higher for workers with low earnings and those without employer-provided health insurance.

The award of DI benefits depends on the following conditions: (1) An individual must file an application; (2) There is a work requirement on the number of quarters of prior employment: Workers over the age of 31 are disability-insured if they have 20 quarters of coverage during the previous 40 quarters; (3) There is a statutory five-month waiting period out of the labor force from the onset of disability before an application will be processed; (4) individuals who work must earn no more than a so-called "substantial gainful amount" (SGA, $1,170 a month for non-blind individuals as of 2017); and (5) the individual must meet a medical requirement, i.e. the presence of a disability. Since this last requirement is the same as in the SSI program, we discuss it below after a short description of SSI.

2.2 The SSI Program

Working-age individuals who are deemed to be disabled and have limited income and limited resources are eligible to receive supplemental security income (SSI).[11] The definition of disability in the SSI program is identical to the one for the DI program, while the definitions of low income and low resources is similar to the one used for the Food Stamps (SNAP) program.[12] SSI benefits are adjusted annually. In 2017, an individual (couple) with no countable income would receive $735 ($1,103) in cash benefits a month.

2.3 The Disability Determination Process

The disability determination process is common to both DI and SSI applicants and consists of sequential steps. Applicants submit their application to a local field office. The case is evaluated by a Disability Determination Service (DDS) officer. There are 4 steps to the evaluation which can be divided into two broad parts: first there are two health evaluation steps; then there are two economic opportunity evaluation steps. The health part is to determine whether the applicant has a medical disability that is severe and persistent. This is 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*

[11]The SSI program serves also children with disabilities and seniors with limited-means (with or without a disability), two groups that are not our focus.

[12]In particular, individuals must have income below a "countable income limit", which typically is slightly below the official poverty line (Daly and Burkhauser, 2003). SSI eligibility also has an asset limit ($2,000 for individuals and $3,000 for couples.).

in death, or which has lasted, or can be expected to last, for a continuous period of at least 12 months. If such disability is a "listed impairment" the individual is awarded benefits without further review.[13] If the applicant's disability does not match a listed impairment, the DDS evaluators try to determine the applicant's residual functional capacity. The second part of the evaluation process assesses economic opportunities in light of the medical determination. Step 3 tries to verify if the individual retains functional capacity for his/her *past* work; and in the last step, step 4, if there is functional capacity for *any* work that would befit the applicant's age, education and general skills.

Only about 37% of applicants are awarded benefits at the initial DDS stage. But rejection can be appealed and about 1/3 of denied applicants do so. The application, which is not updated with new information, is transferred to a different officer within DDS, a stage that is called "reconsideration". The success rate at this stage is even lower than at the initial stage (14%). Those denied at the reconsideration stage can further appeal (and 3/4 of those denied on first appeal do so). These appeals are decided outside of the DDS, by Administrative Law Judges (ALJ), where applications tend to have a much larger success rate (63%).[14] Rarely do cases go beyond the ALJ stage, and if they do the award rates are significantly lower.

3 Theoretical Framework for Type I Errors

The disability insurance process and the decisions by both individuals and the SSA can introduce differences in error rates across gender. Higher Type I errors for woman could in principle arise in various ways: first, women may have a lower threshold for labeling a work-limitation as severe, and this would generate more applications for less objectively work-limited women and hence larger Type I errors; second, women's work limitations may be objectively less severe; third, women may have a lower cost of applying. Finally, from the SSA supply side, women may face tougher standards set by SSA or exhibit noisier signals about the extent of their work limitation than men. In this section, we set-up a simple theoretical framework for distinguishing between these explanations.

Suppose that the true, latent work limitation status of an individual i is given by:

[13]The listed impairments are described in a blue-book published and updated periodically by the SSA ("Disability Evaluation under Social Security"). They are physical and mental conditions for which specific disability approval criteria has been set forth or listed (for example, Amputation of both hands, Heart transplant, or Leukemia).

[14]The higher success rate at this stage partly reflects applicants' self-selection, partly the possibility of integrating the file with new information, and partly the possibility to advocate one's case in court.

$$L_i^* = \alpha_0 + \alpha_L F_i + \varepsilon_i \tag{1}$$

where F_i is a dummy for being a woman and ε_i represents unobserved heterogeneity in work limitations. We omit the contribution of observable characteristics (besides gender) from the equations in this section for simplicity, but fully account for them in the empirical analysis. The female dummy captures potential shifts in the underlying distribution of work limitations: women may have less severe ($\alpha_L < 0$) or more severe ($\alpha_L > 0$) underlying work impairments than men.

Assume that individuals report to be work limited if their latent work limitation status is above a certain threshold, $\bar{L}_i$:

$$L_i = \mathbb{1}\{L_i^* > \bar{L}_i\} \tag{2}$$

An important source of heterogeneity is that men and women may differ in their assessment of the threshold, i.e.,

$$\bar{L}_i = \gamma_0 + \gamma_{\bar{L}} F_i \tag{3}$$

If women have a lower "pain threshold", $\gamma_{\bar{L}} < 0$, then more women than men will classify themselves as work-limited despite their underlying work-limitation being the same, which is the problem of interpersonal comparison of self-reports of work disabilities.

We capture differences in decisions to apply for disability insurance by assuming that people apply if their latent, true work-limitation status exceeds a person-specific threshold:

$$A_i = \mathbb{1}\{L_i^* > \bar{A}_i\} \tag{4}$$

The application threshold $\bar{A}_i$ may differ from the "perceived work-limitation" threshold $\bar{L}_i$ for a number of reasons, although we generally expect them to be positively correlated. For example, applicants may "cheat", i.e., apply even when they are not truly work-limited. Or applicants may have imperfect information about SSA norms, face different transaction costs of applying, or respond differently to financial incentives to work. Further, some individuals may have a very high application threshold if they continue to be productive in the labor market despite the presence of a genuine work-limitation.

We assume that the application threshold differs from the disability threshold by a linear function of characteristics, including gender:

$$\bar{A}_i = \bar{L}_i + \delta_0 + \delta_{\bar{A}} F_i \tag{5}$$

This captures the possibility that men and women differ in their cost of applying as well as in their knowledge of SSA norms, attitudes toward cheating, or sensitivity to financial incentives to work (as documented by Kostøl and Mogstad, 2014). Equations (1)-(5) capture the "demand" side of disability insurance.

The final part of the model that can lead to differences in error rates is in the supply of disability insurance. We assume that SSA assesses work limitation based on a signal S_i^*, such that the award of benefits occur if the signal crosses a threshold:

$$DI_i = \mathbb{1}\{S_i^* > \bar{L}_{SSA}\} \tag{6}$$

To allow for random noise in the signal as well as the possibility of gender-bias in the disability insurance assessment, we assume that the work-limitation signal formed by SSA is given by:[15]

$$S_i^* = L_i^* + \theta_{SSA} F_i + \zeta_i \tag{7}$$

where ζ_i is the signal noise which may be gender-specific. For example, women may be able to provide better documentation about their work limitations than men.

We use this simple framework for the demand and supply of disability insurance to ask which elements of the model shift the probability of Type I errors, and can thus contribute to explaining differences in Type I errors by gender. In Appendix B we use a normality assumption on the distribution of the unobservables to show that the probability of Type I error is higher for women if :

1. Women have less severe work limitations ($\alpha_L < 0$);
2. Women have a lower threshold for reporting a work-limitation ($\gamma_{\bar{L}} < 0$);
3. Women have a lower opportunity cost of applying ($\delta_{\bar{A}} < 0$);
4. The SSA assesses women more strictly than men ($\theta_{SSA} < 0$).
5. The SSA receives a less precise signal for women than for men (the variance of ζ_i is higher for women than for men).

[15] This representation, where the signal observed by the SSA has a gender specific mean-shift, is equivalent to the SSA threshold $\bar{L}_{SSA}$ being gender specific.

Observed differences in Type I errors by gender will reflect the combined effect of these different channels. In Section 6 we use data on various aspects of the disability insurance process, together with disability vignette data, to test which of these various forces are at play.

4 Data

4.1 The Health and Retirement Study (HRS)

The Health and Retirement Study (HRS) is a panel data set administered by the Institute for Social Research at the University of Michigan. Its population target consists of household heads aged 50 and more. We merge a harmonized version of the HRS that has been assembled by the RAND Center for the Study of Aging, containing biannual waves 1992 through 2014, with other HRS data from the raw files. The most relevant variables in this dataset are: (a) the self-reported presence of a work limitation, defined as "an impairment or health problem that limits the kind or amount of paid work" that a respondent can do, together with information about whether the condition is temporary, and whether it prevents work altogether; (b) indicators for the presence of specific health conditions (high blood pressure, diabetes, cancer, lung disease, heart disease, stroke, psychiatric problems, and arthritis), as reported to the respondent by his/her own physician, as well as a variety of other health indicators; and (c) Disability Vignette data (available in a special module of the 2007 wave).

4.2 Social Security Administrative Data

For consenting respondents (approximately 80% of the entire sample), HRS data can be linked to administrative data on earnings and benefits available from the Social Security Administration (the Master Earnings File (MEF), and the Master Beneficiary Record (MBR) file), and to Form 831 Disability Records (F831), which contain information on the initial medical determination (i.e., the outcome of the initial review and of the reconsideration, both done at the SSA level) of an applications to DI and/or SSI. The F831 database does not contain information on decisions made at the ALJ level and beyond.[16] [17] The F831

[16] Also, no F831 case is open if the applicant receives a "technical denial" (i.e., people with earnings above SGA).

[17] In principle, the Master Beneficiary Record (MBR) file can be used to verify whether a DI application was eventually successful by checking whether an individual is receiving social security benefits classified as:

database includes multiple records per individual. We distinguish between application cycles and application rounds. An application cycle may include up to two rounds: the initial DDS assessment, and the DDS reconsideration (if there is one). For each cycle we observe five key variables: (a) the exact application date of any round; (b) the outcome of each application round, together with the exact decision date; (c) the primary impairment (body system) code;[18] (d) the stage at which the application is denied (or awarded); and (e) whether it is a DI, an SSI, or a concurrent DI/SSI application.

4.3 Measures of Disability

To estimate Type I and Type II errors, we need a measure of the "true" work disability status of an individual. As mentioned above, the SSA defines work disability as: "The 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." We replicate this definition using three survey questions from the HRS: (1) "Do you have any impairment or health problem that limits the kind or amount of paid work you could do?"; (2) "Is this a temporary condition that will last for less than three months?"; and (3) "Does this limitation keep you from working altogether?". We classify as disabled people who answer "Yes" to the first and third question and that report that the condition is not temporary. This way, we match very closely the three criteria set forth by the SSA definition: the presence of a work-related impairment, its severity, and its expected duration.

4.4 Descriptive Statistics

Our main estimation sample consists of HRS (non-proxy) respondents who apply for DI/SSI and whose disability status is observed around the time of the application. In principle, one would like to observe the disability status exactly at the time of the application.

"Benefits to a disabled worker". However, there are significant issues with using this information, including non-random selection, left-censoring due to death, transition into OASI and the fact that a case can still be in-progress when the survey ends. Moreover, the Master Beneficiary Record (MBR) file contains no information on SSI receipt.

[18] These are: Musculoskeletal system, Respiratory system, Cardiovascular system, Digestive system, Genito-urinary system, Neurological, Mental disorders, Endocrine system, Multiple body systems, Neoplastic diseases, Immune deficiency, Hemic and lymphatic system, Skin, Growth impairment, Special senses and speech, and Other. We also observe a more detailed sub-categorization (impairment codes) (i.e., for those applying with a Musculoskeletal system body system code, we observe whether it is Disorders of Back (discogenic and degenerative), Osteoarthrosis and Allied Disorders, and so forth). However, the sample sizes are very small and we do not use this information.

Unfortunately, if we were to match only those whose interview date coincides with the date of disability insurance application, we would be left with an extremely reduced sample (especially because HRS is conducted every other year). Instead, we use all applications that we can match with an HRS interview that is no more than 12 months after the application date. To make sure that this criterion is not responsible for our results, we perform several robustness checks.[19]

In Table 1 we report descriptive statistics for the matched sample, comprising 918 first-round applications.[20] Of these applicants, about half of the individuals report not having a work disability. Two comments are in order. First, this is hard to reconcile with a "rationalization" story and more likely to be consistent with the idea that people report truthfully their health conditions to HRS interviewers. Second, our definition of disability which requires people to be completely unable to work is possibly more stringent than the SSA definition, where people can actually work up to the SGA amount. Indeed, if we adopt a weaker definition of disability to allow those who say their work limitation does not stop them working altogether, the fraction of applicants to DI who have a work limitation rises to about 80%. It is therefore even more surprising to find the high rejection rates and Type I error rates we do find.

The denial rate in the sample, shown in Table 1 reproduces almost identically the denial rate observed in the population of all applicants at the initial consideration stage (63%). The (cumulative) denial rate is slightly lower if we also consider the reconsideration stage (58%). In the raw data, there are large differences in the denial rate between women and men (a 12-14 percentage point difference). However, this on its own does not indicate a gender difference in classification errors.

While denial rates at initial consideration are high, they differ substantially by primary disability code and also by gender, as shown in Figure 1. Denial rates are higher for conditions that are harder to verify, such as muscoloskeletal disorders and mental disorders. Denial is

[19] If we include self-reports of work disability that happen much earlier than the interview date we may miss disability insurance applications in response to severe shocks, which is key. To check that our criterion is not generating mis-classifications, we perform the following exercise. We first construct a variable that measures the distance between interview and application date, d. We then compute the fraction of respondent who report to be disabled for each value of $|d| \leq 12$. The fraction is around 20% for $-12 \leq d \leq -2$, jumps discontinuously at $d = -2$ (to about 45%), and remains approximately around 50% for $-2 \leq d \leq 12$ (see Figure A.1 in the Appendix). In Table 6, we report results using the $-2 \leq d \leq 12$ sample and show that they are similar to the baseline.

[20] There are 357 from men and 561 from women. This is partly because the HRS sampling is heads older than 50. Since heads are more likely to be men and wives are younger, we end up with more women "at risk of applying for DI" (e.g., younger than 65) than men.

Table 1: Descriptive statistics

	Men		*Women*		*All*	
	Mean	SD	Mean	SD	Mean	SD
Denial, appl. round	0.54	0.50	0.68	0.47	0.63	0.48
Type I error, appl. round	0.39	0.49	0.63	0.48	0.55	0.50
Type II error, appl. round	0.32	0.47	0.26	0.44	0.29	0.45
Denial, appl. cycle	0.51	0.50	0.63	0.48	0.58	0.49
Type I error, appl. cycle	0.35	0.48	0.57	0.50	0.49	0.50
Type II error, appl. cycle	0.35	0.48	0.29	0.45	0.31	0.46
Disabled	0.47	0.50	0.54	0.50	0.52	0.50
Applied SSI only	0.20	0.40	0.25	0.44	0.24	0.42
Applied DI + SSI	0.19	0.39	0.18	0.38	0.18	0.39
College degree	0.33	0.47	0.28	0.45	0.30	0.46
Black	0.27	0.44	0.30	0.46	0.28	0.45
Married	0.61	0.49	0.47	0.50	0.53	0.50
Widowed	0.04	0.20	0.12	0.33	0.09	0.29
Lab. mark. experience	20.40	8.90	17.15	8.13	18.41	8.58
Age	57.26	4.60	55.72	5.94	56.32	5.51
Type of condition in F831						
Musculoskeletal	0.39	0.49	0.42	0.49	0.41	0.49
Respiratory	0.04	0.19	0.07	0.25	0.06	0.23
Cardiov.	0.20	0.40	0.11	0.31	0.14	0.35
Endocrine	0.05	0.22	0.06	0.25	0.06	0.24
Neurol.	0.08	0.27	0.07	0.26	0.08	0.26
Mental dis.	0.08	0.28	0.10	0.30	0.10	0.29
Cancer	0.03	0.17	0.04	0.19	0.04	0.19
Immune def.	0.03	0.18	0.02	0.16	0.03	0.17
Dig. & Urin.	0.01	0.11	0.03	0.17	0.02	0.15
Other	0.07	0.26	0.08	0.27	0.08	0.26
Number of obs.	357		561		918	

Note: The sample is HRS respondents that are observed in the F831 dataset in their first-round applications (across all application cycles). Respondents are defined as "Disabled" if they report to have an impairment or health problem that limits the kind or amount of paid work they can do; if the condition is not temporary (i.e., lasting less than three months); and if the limitation keeps them from working altogether. Labor market experience is number of years with positive earnings (from the MEF dataset).

less likely for applicants with cancers or disorders of the genito-urinary system (such as kidney failures). This means that when we try to explain differences in denial rates conditioning on self-reported work-limitations, it is crucial to account for the underlying disability conditions individuals are applying for. Higher denial rates and Type I errors among women could be due to the fact that women are more likely to apply with difficult to verify and high denial-rate conditions, such as muscoloskeletal disorders. On the other hand, Figure 1 shows that across all conditions, including those that should be relatively easier to verify - such as cancer - women suffer larger denial rates relative to men.

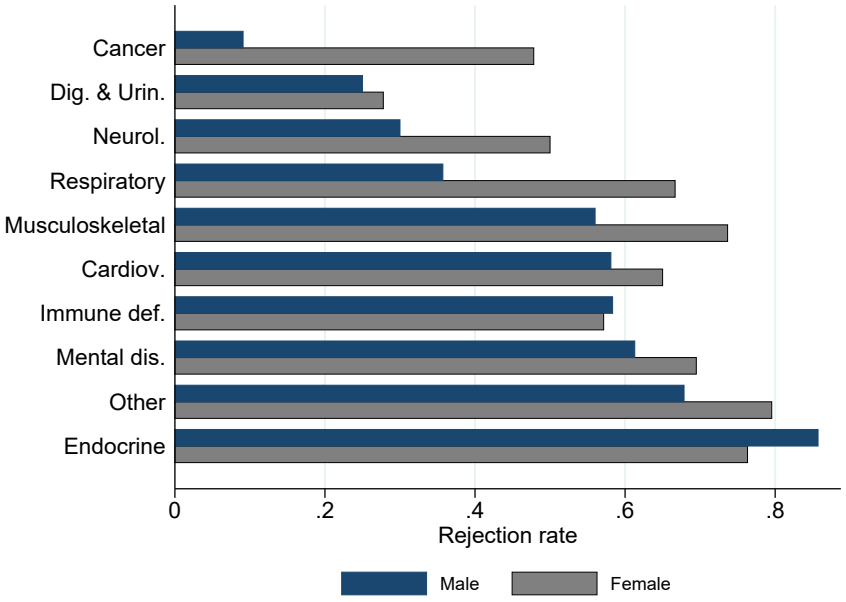

Figure 1: Denial rates by primary disability code and gender

In the raw data reported in Table 1, the difference in Type I errors between men and women is 22-24 percentage points. Some of this could originate from differences in observable characteristics, something that our formal regressions below are designed to account for. Indeed, as Table 1 shows, men and women differ in many important dimensions. Male applicants are older and with more labor market experience, they are more likely to be college-educated, and less likely to be black, unmarried or widowed. Summary statistics for the primary disability condition reported on the F831 application form are in the lower part of Table 1. Men are more likely to apply for disability insurance because of a cardiovascular condition, while women are slightly more likely to apply because of a muscoloskeletal, mental

or respiratory disorder.

4.5 Self-reported disability indicators vs. clinical health measures

Estimation of Type I errors hinge crucially on the reliability of our disability variable in measuring the true underlying work limitation of an individual. The literature has not reached a full consensus on this issue. Benitez-Silva et al. (2004) argue that "...self-reported measures give individuals latitude to summarize a much greater amount of information about [the applicant's] health and disabilities than can be captured in the more objective, but very specific indices". However, as discussed in Bound and Burkhauser (1999), the use of self-reported disability measure raises two basic issues: (a) endogeneity with respect to labor market outcomes (i.e., those who apply for DI or are out-of-work are more likely to self-report a disability as a way of rationalizing their decisions), and (b) inter-personal comparability.

We address these concerns directly. Regarding the first issue, we can verify whether self-reported disability is associated with more objective or clinical indicators of disability for which there is less scope for rationalization. This is what we do in Table 2 below. Regarding the second issue, we use responses to disability vignettes to generate within person responses, as we argue in section 6.

The HRS contains rich information on the health of respondents which are of a more objective or diagnostic nature. First, respondents are asked whether they have difficulties with basic activities in their daily living (ADL's), such as dressing, preparing meals, etc., because of a health condition.[21] Second, we observe some objective indicators of poor health, such as whether a person has spent some time in hospital and for how long, BMI data (so we can determine obesity or being underweight), and whether people leave the sample because of death. Finally, we have information on whether a doctor has told the respondents that they have some specific condition, like high blood pressure, cancer, etc.

Table 2 compares average values of these various health indicators, splitting the sample by self-reports into the "disabled" and "not disabled" separately for men and for women. Clearly, people who self-report a disability are much more likely to have a clinical or diagnostic health condition, and more likely to encounter difficulty in ADL's. For example,

[21] The presence of ADL difficulties plays an important role in the official determination of disability. For example, many DI/SSI applicants are required to fill in an "Activities of Daily Living Form" report (known as the Function Report, SSA-3373). Moreover, long-term care insurance policies require that an applicant needs help with two or more ADL before triggering benefits.

Table 2: Health conditions by self-reported work limitation status and gender

	Women			*Men*		
	Not disabled	*Disabled*	*Diff. (cond. on age)*	*Not disabled*	*Disabled*	*Diff. (cond. on age)*
Difficulty walking	0.0279	0.2142	0.170***	0.0189	0.1712	0.150***
Difficulty dressing	0.0401	0.2301	0.177***	0.0428	0.2315	0.198***
Difficulty stooping, etc.	0.3816	0.8156	0.381***	0.2952	0.7426	0.386***
Difficulty getting out of bed	0.0421	0.2520	0.162***	0.0317	0.2222	0.152***
Difficulty grocery shopping	0.0442	0.3143	0.248***	0.0269	0.2070	0.189***
Difficulty preparing meals	0.0229	0.1735	0.148***	0.0151	0.0975	0.109***
Hospital stay	0.1535	0.4106	0.229***	0.1637	0.4439	0.249***
Nights in hospital	1.0253	4.9840	3.441***	1.2312	6.9819	4.783***
Obese	0.3226	0.4530	0.116***	0.2972	0.3186	0.034***
Underweight	0.0129	0.0230	0.004***	0.0034	0.0180	0.011***
Died in sample	0.2132	0.4011	0.091***	0.3023	0.5110	0.128***
Doctor diagnosed HBP	0.3980	0.6252	0.171***	0.4311	0.6342	0.148***
... psychological condition	0.1807	0.4194	0.182***	0.1020	0.2817	0.133***
... heart condition	0.1057	0.2961	0.175***	0.1467	0.3916	0.212***
... arthritis	0.4735	0.7661	0.217***	0.3629	0.6068	0.193***
... diabetes	0.1198	0.2671	0.126***	0.1450	0.2883	0.127***
... lung condition	0.0726	0.2240	0.129***	0.0546	0.1806	0.133***
... stroke	0.0260	0.1039	0.076***	0.0330	0.1349	0.100***
... cancer	0.0819	0.1168	0.040***	0.0528	0.1026	0.049***

Note: The unit of observation is a person-HRS wave for all variables except death, where it is just person. Respondents are defined as "Disabled" if they report to have an impairment or health problem that limits the kind or amount of paid work they can do; if the condition is not temporary (i.e., lasting less than three months); and if the limitation keeps them from working altogether. In the third and sixth columns we use regression analysis and report the marginal effect of the dummy for being disabled on the row variable (controlling for age). *** means significance at 1 percent level (s.e. clustered at the individual level). The sample is individuals aged 20-65 only.

only about 3% of not disabled women have trouble walking across a room, as opposed to 20% in the disabled group. There are similarly large differences for other ADLs. Mortality for women is 21% vs 40%. Hospital stays are almost three times more likely and five times longer among the disabled group. Finally, the disabled are much more likely to have been diagnosed with a serious health condition. Results for men display very similar quantitative evidence. One concern with unconditional comparisons is that they may just reflect the fact that older people are more likely to be disabled and in poor health. In columns (3) and (6) of Table 2 we report the coefficient on the "Disabled" dummy for each one of the row variables, while controlling for the age of the respondent. The differences are attenuated, but still very sizable and statistically significant throughout. For example, controlling for age, women who self-report to be disabled have a 9 percentage point higher probability of dying within the period covered by the survey than women who report not to be disabled.

5 Gender Differences in Classification Errors

5.1 Baseline Estimates

We estimate the effect of gender on Type I errors by running the following probit model for applicants who report to be work limited ($L_{ij} = 1$):

$$\Pr(Reject_{ij}|L_{ij} = 1) = \Phi(X'_{ij}\psi_0 + \psi_1 F_i) \tag{8}$$

For Type II errors, we run the following probit model for applicants who do not report a work limitation ($L_{ij} = 0$):

$$\Pr(Award_{ij}|L_{ij} = 0) = \Phi(X'_{ij}\kappa_0 + \kappa_1 F_i) \tag{9}$$

where i is individual, j is application, X_{ij} includes individual- and application-level controls, and F_i is a female dummy. Our primary focus is on outcomes at the initial consideration stage. This is the least problematic stage, since award at reconsideration and further stages are affected by various forms of selection.

The first four columns of Table 3 report results for Type I errors; the last four columns focus on Type II errors. Columns (1) and (5) reproduce the unconditional difference noted in Table 1, while the other columns add a variety of socio-demographic and health characteristics of applicants. We find statistically significant higher Type I error rates for women: a 20.8 percentage point difference in the richest specification of column (4). Older applicants are

less likely to be turned down if truly disabled.[22] Occupation dummies are jointly statistically significant. The importance of occupational controls is that rejection could be greater for those in occupations where retaining some functional capacity is more likely (i.e., a sedentary job).[23]

Table 3: Probit regressions for Type I and Type II errors

	Type I error				*Type II error*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	0.229***	0.205***	0.215***	0.208***	-0.064	-0.028	-0.057	-0.069
	(0.045)	(0.047)	(0.054)	(0.046)	(0.045)	(0.045)	(0.041)	(0.045)
College degree		-0.021	-0.027	-0.035		0.037	0.001	-0.007
		(0.055)	(0.054)	(0.053)		(0.048)	(0.046)	(0.048)
Black		-0.031	-0.014	0.014		-0.025	-0.027	-0.063
		(0.059)	(0.056)	(0.058)		(0.047)	(0.046)	(0.046)
Lab. mark. exp.		-0.007**	-0.001	-0.005		0.005*	-0.000	0.001
		(0.003)	(0.004)	(0.004)		(0.003)	(0.003)	(0.003)
Applied SSI only		-0.086	-0.083	-0.078		0.114**	0.140**	0.147***
		(0.059)	(0.056)	(0.058)		(0.055)	(0.055)	(0.054)
Applied DI + SSI		0.051	-0.006	0.016		0.029	0.055	0.079
		(0.065)	(0.064)	(0.064)		(0.059)	(0.057)	(0.055)
Married		0.056	0.032	0.052		0.024	0.075	0.076*
		(0.055)	(0.050)	(0.054)		(0.049)	(0.046)	(0.045)
Widowed		-0.005	-0.065	-0.019		0.020	0.029	0.050
		(0.088)	(0.081)	(0.081)		(0.082)	(0.086)	(0.084)
Age		-0.014***	-0.015***	-0.016***		0.015***	0.016***	0.020***
		(0.005)	(0.005)	(0.005)		(0.005)	(0.005)	(0.005)
Year FE	No	No	Yes	Yes	No	No	Yes	Yes
F831 disab. FE	No	No	Yes	Yes	No	No	Yes	Yes
HRS Obj. FE	No	No	No	Yes	No	No	No	Yes
ADL FE	No	No	No	Yes	No	No	No	Yes
BMI+Hosp	No	No	No	Yes	No	No	No	Yes
Occupation FE	No	No	No	Yes	No	No	No	Yes
[P-value joint sig.]				[0.001]				[0.004]
Observations	473	473	473	447	445	445	421	411

Note: Standard errors in parentheses, clustered at the individual level. The reported coefficients are marginal effects. Labor market experience is number of years with positive earnings (from the MEF dataset). *** means significance at 1 percent level.

[22]The age effect may be non-linear since the vocational grid changes the eligibility rules at ages 50, 55 and 60 (see Chen and van der Klaauw, 2008). Column (1) of Table A.1 reports a finer specification with an age spline at these knots. The results are unchanged.

[23]We use 18 occupational codes. The results are similar if we replace these dummies with a physical occupational requirement index using a mapping between HRS occupational codes and O*NET data (as in Michaud and Wiczer, 2018). See column (2) of Table A.1 in the Appendix.

The negative coefficient on the female dummy in columns (5)-(8) of Table 3 shows that the effect of gender on Type II error is consistent with the idea that women applicants are "less believed", both when they are truly disabled and when they are not severely disabled. However, the estimates are statistically insignificant. For this reason, from now on we will focus on the evidence for Type I error. The focus on Type I error is also because of our interest in the effectiveness of insurance aspects (as opposed to the moral hazard aspects) of disability insurance.

5.2 Type I Errors at Different Stages of the Evaluation Process

As discussed in Section 2, rejection of an application can take place for different reasons: a "medical" rejection because the health condition is assessed as not being severe and long-lasting; or a rejection on "vocational" grounds because there is residual functional capacity to return to work, either at the previous job or at a different job. The data from the SSA specifies at what stage the application is turned down and we use this data to establish at which stage in the decision process the gender difference in Type I errors arises.

Table 4 breaks down the decision into rejections on health criteria alone ("medical stage") and rejections on the basis of availability of work ("vocational stage"). The coefficient on being a woman is insignificant for the medical stage indicating there is no difference between men and women in the way health is assessed. However, among those who report having a work-limitation, women are 18.7 percentage points more likely than men to be rejected at the vocational stage.

The clear message is that the difference between men and women in Type I errors arises because of the SSA evaluator's different expectations about the applicant's ability to perform previous or other work that befits their skills, experience, and age. This difference is present despite controlling for occupational dummies and other characteristics: men and women are assessed to have systematic differences in their residual capacity despite being observationally equivalent in many dimensions.

Table 4: Type I errors: Rejection at medical or vocational stage

	Medical stage	*Vocational stage*
	(1)	(2)
Female	0.055	0.187***
	(0.043)	(0.057)
College degree	0.032	-0.008
	(0.044)	(0.060)
Black	-0.088*	0.007
	(0.046)	(0.058)
Lab. mark. exp.	0.001	-0.000
	(0.003)	(0.004)
Applied SSI only	0.087**	-0.121*
	(0.043)	(0.064)
Applied DI + SSI	0.091**	-0.011
	(0.045)	(0.066)
Married	0.035	0.042
	(0.042)	(0.055)
Widowed	0.036	-0.033
	(0.067)	(0.092)
Age	-0.000	-0.019***
	(0.004)	(0.005)
Year FE	Yes	Yes
F831 disab. codes FE	Yes	Yes
HRS objective FE	Yes	Yes
Health cond. FE	Yes	Yes
ADL FE	Yes	Yes
BMI+Hosp	Yes	Yes
Occupation FE	Yes	Yes
Sample average	0.153	0.442
Observations	380	389

Note: Standard errors in parentheses, clustered at the individual level. The reported coefficients are marginal effects. Labor market experience is number of years with positive earnings (from the MEF dataset). ***, **, and * means significance at 1, 5 and 10 percent level, respectively. In column (1) the outcome variable equals 1 if rejection is due to not meeting "medical criteria" (i.e., impairment is deemed not severe or not expected to last 12 months or more). In column (2) the outcome variable equals 1 if rejection is due to not meeting "vocational criteria" (i.e., SSA determines that there is capacity for SGA - past relevant work, or capacity for SGA - other work).

5.3 Implications of Type I Errors

Our conclusion to this point is that the larger Type I errors for women arise not because of different assessments about health, but rather different assessments about their residual functional capacity. In other words, the SSA is assessing women who self-report to be work-limited as being able to go back to work, while they would not assess a man with the same self-reported work limitation as being able to go back to work.

There are two alternative explanations of this difference by the SSA: the first is that the SSA may have extra information in the assessment that is not available to us as econometricians. The form that applicants fill in is much more detailed than we have access to (although we do use the rich health information from the HRS): it asks about daily activities, who takes care of the house, other activities, and so on (this is known as the "Function Report" form, or form SSA-3373-BK). The second is that the SSA has genuinely overestimated the residual functional capacity of women. Support for the first explanation would be if we saw rejected women who self-report being work-limited actually returning to work after rejection for DI. Or more precisely, among rejected women, we would see women with a self-reported limitation returning to work at the same rate as women who do not self-report a limitation. Evidence for the second explanation would be the opposite: we would see significant differences in labour market outcomes between the two categories of rejected women applicants.

Column (1) of Table 5 reports the results of this test where our dependent variable is whether the individual had any employment in the three years following the initial consideration stage.[24] We define employment in a given year as the individual having SSA earnings above the SGA amount for that year: this means no individuals who may have moved onto disability insurance following appeal can be classified as employed. Since we want to compare rejected applicants with and without a self-reported work limitation, we regress on a dummy for being rejected and on the interaction of the latter with reporting a work-limitation.

Among women whose DI application is rejected, those with self-reported work-limitations are (unconditionally) much less likely to be earning above SGA amounts (2 percent vs. 19 percent - see the estimated proportions reported at the bottom of the table). This difference remains unchanged, and is highly statistically significant, once we control for health and socio-demographics characteristics. We interpret this as evidence that the apparent high

[24]The results are similar if we look at five year post-decision outcomes.

Type I errors for women reflect true Type I errors: the SSA has overestimated the residual functional capacity of women who self-report work-limitations.

Table 5: Impact of DI Decision on Subsequent Work

	1-3 yrs after		*5-10 yrs before*
	Women	Men	Women
	(1)	(2)	(3)
Rejected	0.175***	0.133*	0.019
	(0.040)	(0.049)	(0.044)
Rejected × Disabled	-0.168***	-0.020	0.014
	(0.036)	(0.061)	(0.044)
Other controls	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
F831 disab. codes FE	Yes	Yes	Yes
HRS objective FE	Yes	Yes	Yes
Health cond. FE	Yes	Yes	Yes
ADL FE	Yes	Yes	Yes
BMI+Hosp	Yes	Yes	Yes
Occupation FE	Yes	Yes	Yes
Proportions working			
$\{R = 0, L = 0\}$	0.02	0.04	0.75
$\{R = 0, L = 1\}$	0.04	0.04	0.76
$\{R = 1, L = 0\}$	0.19	0.18	0.77
$\{R = 1, L = 1\}$	0.02	0.12	0.77
Observations	435	300	430

Note: Standard errors in parentheses, clustered at the individual level. Dependent variable is employment, defined as earning at least as much as the SGA. The indicators R and L equal one if the individual's disability insurance application has been rejected and if the individual is disabled, respectively. Respondents are defined as "Disabled" if they report to have an impairment or health problem that limits the kind or amount of paid work they can do; if the condition is not temporary (i.e., lasting less than three months); and if the limitation keeps them from working altogether.

Columns (2) and (3) of Table 5 report two different placebo tests. Column (2) reports the same regression for men. Among men who are rejected for DI, there is no statistically significant difference in the rate at which they return to work by whether they are work-limited. This suggests the SSA has not made an error in assessing residual functional capacity of men who claim to be work-limited.

Column (3) reports the regression with the dependent variable being whether the individual woman was working 5-10 years *prior* to the DI decision. This is intended as a placebo test to rule out the presence of unobserved heterogeneity among women driving our results. Unconditionally, there are no differences in the probability of working, and the proportions appear independent of future reports of work limitation or disability insurance application outcome. This is confirmed in the formal regression that controls for respondent characteristics.

5.4 Robustness and Extensions

We perform various robustness checks of the key finding that women experience larger Type I errors than men. These extra results are presented in Table 6. For comparison, column (1) reproduces the results of the baseline specification (from Table 3, column (3)). All regressions include the same controls used in the baseline specification.

Table 6: Probit regressions for Type I errors: Robustness

	Basel.	*DI Sample*	*Timing assumptions*			*Different disab. def.*	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
			$-2 \leq d \leq 12$	$0 \leq d \leq 9$	$0 \leq d \leq 12$, weighted	Less strict. disab. def.	At least two ADL's
Female	0.215***	0.262***	0.192***	0.163**	0.193***	0.118***	0.169***
	(0.054)	(0.059)	(0.054)	(0.065)	(0.055)	(0.040)	(0.058)
Sample avg.	0.55	0.54	0.54	0.54	0.55	0.60	0.57
Obs.	447	335	499	339	447	764	325

Note: Standard errors in parentheses, clustered at the individual level. The reported coefficients are marginal effects. All regressions include the controls of Table 3, column (2). ***, **, and * means significance at 1, 5 and 10 percent level, respectively. In columns (1)-(5) respondents are defined as "Disabled" if they report to have an impairment or health problem that limits the kind or amount of paid work they can do; if the condition is not temporary (i.e., lasting less than three months); and if the limitation keeps them from working altogether. In column (6) a disabled respondent is one who reports to have an impairment or health problem that limits the kind or amount of paid work he/she can do. In column (7) respondents are defined as "Disabled" if they report to have difficulty with at least two ADL's.

A first concern is that higher rejection rates for women are due to inherent gender bias against welfare program recipients (as the "welfare queens" literature in sociology has remarked, Hancock, 2004). To address this issue, we drop SSI applications and zoom in on the DI sample, which is also the program more traditionally studied in the literature. If we focus only on DI applicants, the results are confirmed, and if anything there is a larger gender difference.

Our baseline sample includes individuals who are interviewed within 12 months from the date of their disability insurance application. Since the "timing" of the match is arbitrary, in columns (3)-(5) we experiment with different assumptions. In column (3) we use those interviewed 2 months before to 12 months after the application date. In column (4) we focus on those interviewed up to 9 months following the date of application. Finally, in column (5) we use the same criterion of the baseline, but weight more those interviewed closer to the application date (we use as weight $1/\sqrt{d}$, where d is the distance between the date of the interview and the date of the disability insurance application). If people recover from a disability, this criterion is the closest we can get to the "true" disability status at the point of application. While the results change slightly (ranging from 0.16 to 0.19), they are qualitatively similar to the baseline and not significantly different: among applicants, women experience higher Type I errors than observationally equivalent men.

A third concern is about our definition of work-limitation, which may capture the true disability status of an individual only imperfectly. In the final columns (6) and (7) of Table 6 we adopt different disability definitions. Our baseline definition is that a person has a non-temporary impairment that prevents work altogether, and this may be even stricter than the one adopted by SSA where applicants and recipients are permitted to do some work as long as pay remains below the SGA. In column (6) we classify as disabled those who report to have an impairment or health problem that limits the kind or amount of paid work they can do. This is the standard binary definition of disability used in many papers in the literature and is less strict. In column (7) we assume that an individual is disabled if he/she reports difficulties with two or more activities of daily living. We adopt this definition because it is the one used by Long-Term Care Insurance policies for triggering payment of benefits. The samples are clearly different than the baseline, and yet qualitatively, the estimates are very similar, confirming the presence of significant gender differences in Type I errors.

Another concern is that DDS evaluators may be less lenient towards applicants who have experienced economically-motivated declines in earnings or towards those who are not the primary earners in their household. However, if we add controls for average earnings in the five years preceding application and for being the primary earner in the household (a dummy that equals one if average earnings in the five years preceding application represent more than 50% of household earnings over the same period), the results remain similar.[25]

[25]We also checked if large rejection rates for women are explained by SSA perception of lower economic vulnerability. To check this, we interacted the female dummy with a dummy for being married and with a dummy for having a spouse with positive earnings. Both interactions are insignificant, see column (3) of

A different concern is that DDS evaluators may be less likely to award benefits to women if they believe that women apply with less permanent conditions than men. To check this, we use the whole HRS sample of individuals below age 65 and regress self-reports of disability in wave s against self-reports of disability in wave $s-1$ and interact the latter with a female dummy, while controlling for demographics and other health variables. We find (results reported in the Appendix, Table A.2) that the interaction is statistically insignificant.

Finally, one may wonder whether errors made at the initial evaluation stage are corrected at later stages through the appeal process. Starting with overall DDS experience, we notice that adding a reconsideration stage, where Type I errors may in principle be rectified, does not change the results: women are statistically significantly more likely to be turned down when disabled than observationally equivalent men whether or not we add this first appeal stage. In particular, the female coefficient goes from 0.208 (s.e. 0.046) to 0.201 (s.e. 0.052) when we consider the cumulative success rate at the DDS level (initial stage plus potential reconsideration; see column (4) of Table A.1 in the Appendix). Unfortunately, we do not have data on outcomes of further appeals (ALJ and beyond). But it must be stressed that even if errors were corrected at later stages, there would still be important welfare implications from rejections at the DDS level. DI receipt offers important consumption smoothing benefits, as recently documented by Autor, Kostøl, Mogstad, and Setzler (2019). The welfare costs of going uninsured are compounded by the long time it takes for an application to be processed past the DDS level. A study by the Office of the Inspector General for fiscal year 2006 estimated that the average (cumulative) processing times for a disability insurance application were 131 days for the initial DDS decision, 279 days for the DDS reconsideration decision, 811 days for the ALJ decision, and 1,720 days for a Federal Court decision. Other welfare implications related to time to final decision pertain to human capital. Autor, Maestas, Mullen, and Strand (2015) argues that longer processing times reduce the employment and earnings of DI applicants for multiple years following application, with the effects concentrated among applicants denied benefits at the initial stage.

6 Testing the Mechanisms

The stark conclusion that Type I error rates are greater for women than for men does not on its own provide evidence of bias. Instead, as outlined in the framework of Section 3,

Table A.1 in the Appendix.

there are multiple potential drivers of the gender differences. On the demand for DI side, women may have less underlying work limitations, be more ready to report a severe work limitation, or apply for DI more readily. On the supply side of DI, the SSA may have less informative signals of work limitations for women or have tougher standards. To evaluate these drivers, we carry out a testing strategy based on the sign of the structural parameters (α_L, $\gamma_{\bar{L}}$, $\delta_{\bar{A}}$, θ_{SSA}, and $\sigma^2_\zeta(F)$). We rely on the sign rather than the quantitative magnitudes because the parameters are identified only up to scale because of the latent framework that we adopt.

6.1 Estimation Strategy

Identification of the structural parameters (α_L, $\gamma_{\bar{L}}$, $\delta_{\bar{A}}$, θ_{SSA}, and $\sigma^2_\zeta(F)$) will use information from four margins: (a) self-reports of work limitations, (b) disability insurance applications, (c) disability insurance application outcomes, and (d) disability vignettes. The vignettes are crucial for two reasons: first, they help (under some restrictions) to pin down inter-personal differences in perceptions of work limitations; second, they show potential gender-bias in assessing applicants' work limitations.

A key identification problem is that it is hard to separate actual from perceived work limitations. This identification problem can be seen by considering the case in which we had access only to data on self-reported work limitations. We take the decision to self-report a work limitation from equations (1)-(3) in section 3:

$$L_i = \mathbb{1}\{L_i^* > \bar{L}_i\}$$

Assuming the distribution of the underlying work limitation is normal, $\varepsilon_i \sim N(0,1)$, and adding controls X_i, the probability of self-reporting a limitation is:

$$\Pr(L_i = 1|X_i, F_i) = \Phi((\alpha_0 - \gamma_0) + X_i'\alpha_x + (\alpha_L - \gamma_{\bar{L}})F_i) \tag{10}$$

where $\Phi(.)$ is the CDF of the standard normal distribution. Clearly, from data on self-reports alone we cannot separately identify differences in pain thresholds, $\gamma_{\bar{L}}$, from differences in underlying work limitations, $\alpha_L < 0$. Hence, it is impossible to assess whether women are more likely to suffer larger Type I error because they have a lower pain threshold ($\gamma_{\bar{L}} < 0$) or because their underlying work limitation is less severe ($\alpha_L < 0$).

In addition to data on self-reports of work limitations, we have information on the decision of whether to apply for disability insurance. We take the decision to apply from equations

(4) and (5) in section 3:

$$A_i = \mathbb{1}\{L_i^* > \bar{A}_i\}$$

Adding controls, the probability of applying for disability insurance is then:

$$\Pr(A_i = 1|X_i, F_i) = \Phi((\alpha_0 - \gamma_0 - \delta_0) + X_i'(\alpha_x - \delta_x) + (\alpha_L - \gamma_{\bar{L}} - \delta_{\bar{A}})F_i) \tag{11}$$

Equation (11) shows how we can estimate the shift in the application decision due to gender. For example, if women have worse knowledge of the SSA norms or lower costs of applying, then $\delta_{\bar{A}} < 0$, and they will suffer larger Type I error. By combining estimates from equations (10) and (11), we can identify $\delta_{\bar{A}}$. However, we are still unable to separate α_L from $\gamma_{\bar{L}}$.

Outcomes of DI applications (i.e., whether a claim is rejected) are the unconditional equivalent of the Type I error regressions we considered above. In particular, the SSA decision to reject a claim is based on equations (6) and (7) from section 3:

$$DI_i = \mathbb{1}\{S_i^* > \bar{L}_{SSA}\}$$

Assuming that $\zeta_i \sim N(0, \sigma_\zeta^2(F_i))$ to allow for the variance of the signal to be gender specific, the probability that a claim is rejected can be written as:

$$\Pr(R_i = 1|X_i, F_i) = \Phi((1 + \sigma_\zeta^2(F_i))^{-1/2}((\bar{L}_{SSA} - \alpha_0) - X_i'\alpha_x - (\alpha_L + \theta_{SSA})F_i) \tag{12}$$

Equation (12) adds the key "supply-side" parameters θ_{SSA} and $\sigma_\zeta^2(F_i)$. While $\sigma_\zeta^2(F_i)$ can be identified from the nonlinearity, the parameters α_L, $\gamma_{\bar{L}}$, and θ_{SSA} are still not identified

To make progress on identification, we use disability vignette data available in the 2007 wave of the HRS. The use of disability vignettes has been pioneered in the disability literature by Kapteyn et al. (2007) to separate shifts in the underlying work limitation distribution from subjective evaluation of severity thresholds, precisely the identification problem faced in our context.

In the disability vignette literature, respondents are asked to assess, on the same scale on which they asses themselves, the extent of disability in hypothetical situations and for hypothetical individuals. The 2007 Disability Vignette Survey is a special, mail-only supplement of the HRS.[26] Respondents are first asked if they have a health limiting condition ("Do you have any impairment or health problem that limits the kind or amount of work

[26]The HRS conducted a vignette survey also in 2004 (a "leave behind" supplement), but there was no gender randomization involved, so we focus on the 2007 version.

you can do?"), and to rank it in terms of severity (possible responses are "None", "Mild", "Moderate", "Severe" and "Extreme"). To match the analysis from the first part of the paper we convert these responses into a binary indicator, and assume a person is disabled if he/she answers that the limitation is "Severe" or "Extreme". Next, each respondent is presented with nine vignette scenarios in total, with three scenarios for each health condition ("Depression", "Pain", and "Cardiovascular disease") describing individuals with different degrees of work limitation for that condition. As an example, one of the vignettes reads: "X has pain in [his/her] back and legs, and the pain is present almost all the time. It gets worse while [he/she] is working. Although medication helps, [he/she] feels uncomfortable when moving around, holding and lifting things at work. How much is X limited in the kind or amount of work [he/she] could do?". Possible responses are "None", "Mild", "Moderate", "Severe" and "Extreme", which we again convert into a binary disability indicator if the response is "Severe" or "Extreme". Hence, people are asked to rank the vignettes using the same severity scale that was used to rank their own work limitation. The key aspect is that the order of the vignettes and the gender assigned to the hypothetical person described in the vignettes are randomised. Hence for some people the X person above is a "Mark" and for another respondent the same description refers to a "Tamara".

We assume that respondent i evaluates the latent disability status of vignette v according to the following equation:

$$L^*_{v,i} = \theta_{v,i} + \xi_{v,i} \tag{13}$$

where $\xi_{v,i} \sim N(0,1)$, and classifies the vignette as work-limited if the underlying latent variable crosses a threshold, i.e.:

$$L_{v,i} = \mathbb{1}\{L^*_{v,i} > \bar{L}_{v,i}\} \tag{14}$$

As discussed by Kapteyn et al. (2007), the two key identification assumptions that are standard in this literature are: (1) Vignette Equivalence, and (2) Response Consistency. The first assumption is that the situation described in the vignette is perceived by all respondents in the same way, that is, $\theta_{v,i} = \theta_v$ for all i. This is because all respondents are presented with the identical description of an hypothetical person, and the only differences in their perceptions should be random, such as misreading the sentence. The second assumption is that respondents evaluate the work limitation of the vignette characters in the same way that they evaluate their own, i.e., the threshold they use for the vignette is the same as they would use for themselves: $\bar{L}_{v,i} = \bar{L}_i$ for all i. This assumption means we can use the

information from vignettes to separate out differences in pain thresholds from differences in true work disability. Next, we discuss how the vignettes can be used to consider the presence of gender bias.

In the simple model of Section 3 we allowed for specific gender-bias in disability insurance assessment. In principle, one could estimate the extent of bias by running an audit study in which two identical applications (one by a man, one by a women) are assigned to a DDS evaluator (as in the empirical strategy of Neumark et al., 1996, and Bertrand and Mullainathan, 2004). Any gender differences in award rates could thus be attributed to an inherent gender bias. Unfortunately, this experiment is not feasible. However, one can use the randomization of the vignette's gender in the vignette data to approximate this ideal but infeasible experiment. In particular, enforcing the Vignette Equivalence assumption and adding controls to capture observable heterogeneity in vignette assessment, we rewrite the perceived disability status of a vignette (equation (13)) as:

$$L^*_{v,i} = \theta_v + X_i'\theta_x + \theta_{SSA}F_v + \xi_{v,i} \tag{15}$$

where the key assumption is that the "bias" of HRS respondent in assessing the disability of female vignettes ($F_v = 1$) reproduces the "bias" of DDS evaluators in assessing the disability of female Disability Insurance applicants (captured by the parameter θ_{SSA}).[27] If the respondent sets higher disability standards for women, he/she would be less likely to classify the vignette as disabled if that vignette describes a woman, i.e., $\theta_{SSA} < 0$.

Using the Response Consistency assumption, one can use equation (3) and equation (15) to write the probability that respondent i classifies the vignette as disabled:

$$\Pr(L_{v,i} = 1|X_i, F_v, F_i) = \Phi(\theta_v - \gamma_0 + X_i'\theta_x + \theta_{SSA}F_v - \gamma_{\bar{L}}F_i) \tag{16}$$

This shows that differences in "pain thresholds" by gender ($\gamma_{\bar{L}}$) can be pinned down by how men vs. women respondents differ in their evaluation of the vignette's disability. In practice, the equations (10), (11), (12), and (16) form a system of overidentifying equations for the parameters of interest.

To summarize, the higher Type I errors for woman shown in the reduced-form empirical analysis of section 5 originates from several different channels: (a) women may have a lower pain threshold, $\gamma_{\bar{L}} < 0$ or (b) a lower cost of applying ($\delta_{\bar{A}} < 0$) (which both generate

[27]The obvious caveat is that, unlike the typical HRS respondent, DDS evaluators are professionals. On the other hand, identification of the parameter θ_{SSA} uses also actual denial decisions, equation (12).

more applications from women who are (objectively) less work-limited), (c) women's work limitations may be objectively less severe ($\alpha_L < 0$), (d) women may face tougher admission standards set by SSA ($\theta_{SSA} < 0$) or (e) have noisier signal ($\sigma^2_\zeta(F_i = 1) - \sigma^2_\zeta(F_i = 0) > 0$). To estimate these key parameters, we use restrictions coming from data on self-reported disability (10), disability insurance applications (11), disability insurance claim outcomes (12), and vignettes' disability evaluation (16). Apart from the DI/SSI application outcome regression, the other regressions use the entire HRS sample. Appendix C describes the Minimum Distance procedure we use.

6.2 Parameter Estimates

This section presents probit estimates for the decision to self-report a work-limitation, the decision to apply for DI/SSI, and the probability of having a DI/SSI claim rejected. We next present results of the vignette analysis. Finally, we use these estimates as inputs in a minimum distance framework to pin down the mechanism parameters.

Disability Self-Reports, Application Decisions, and Application Outcomes

Table 7 reports the results of probit regressions for disability self-reports, disability insurance applications, and first round application denials (equations (10), (11), and (12)). In the first case, our outcome variable equals 1 if the individual reports to be disabled (defined as in the baseline regressions) and zero otherwise. In the second case, the outcome variable equals 1 if we observe an "open" first-round application to DI or SSI at any point in time in a given calendar year t for individual i, and 0 otherwise (i.e., an application that is either unadjudicated at the time of the HRS interview or was adjudicated in the same interview year). For the observations with $Applied_{it} = 0$, we focus on those person-years in which applying to DI or SSI is in a person's choice set. We implement this condition by dropping person-years in which $Applied_{it} = 0$ and the individual is a recipient of SSI or DI benefits in that year. The sample only includes people below age 65. For the third case, we focus on the matched HRS/F831 sample we used for the Type I/Type II regressions, but without conditioning on the self-reported work limitation status. In this case, we estimate a heteroskedastic probit model allowing the variance of the error term to depend on the applicant's gender.

In all three regressions, the key variable is the female dummy. We add the same controls used in the Type I regressions above (except the F831 disability code dummies which are

Table 7: Probits for disability self-reports, DI/SSI application, and DI/SSI claim rejections

	(1) *Self-report a disability*	(2) *Applies for disab. insur.*	(3) *DI/SSI claim rejected*
Female	-0.0049	-0.0031*	0.1047***
	(0.0035)	(0.0017)	(0.0396)
College degree	-0.0129***	-0.0063***	-0.0126
	(0.0033)	(0.0018)	(0.0354)
Black	0.0142***	0.0080***	0.0571
	(0.0039)	(0.0019)	(0.0341)
Lab. mark. exp.	-0.0025***	0.0005***	-0.0026
	(0.0002)	(0.0001)	(0.0024)
Married	-0.0081**	-0.0081***	-0.0185
	(0.0033)	(0.0016)	(0.0325)
Widowed	-0.0003	-0.0070**	-0.0849
	(0.0056)	(0.0032)	(0.0525)
Age	0.0012***	-0.0009***	-0.0143***
	(0.0003)	(0.0001)	(0.0038)
λ			-0.62
			(0.38)
Health cond. FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
ADL FE	Yes	Yes	Yes
BMI+Hosp	Yes	Yes	Yes
Occupation FE	Yes	Yes	Yes
Sample average	0.07	0.02	0.63
Observations	40151	42104	914

Note: Standard errors in parentheses, clustered at the individual level. The reported coefficients are marginal effects. ***, **, and * means significance at 1, 5 and 10 percent level, respectively. Labor market experience is number of years with positive earnings (from the MEF dataset). In column (1) respondents are defined as "Disabled" if they report to have an impairment or health problem that limits the kind or amount of paid work they can do; if the condition is not temporary (i.e., lasting less than three months); and if the limitation keeps them from working altogether.

not observed for the whole HRS sample). We find that after controlling for a wide variety of characteristics (especially, health variables), women are *less* likely to report a work limitation and less likely to be applicants, although the marginal effects are small and imprecise. The fact that women are less likely to apply for disability insurance than men, given self-reported work limitation, is also hard to reconcile with the idea that women are more likely than men to "rationalize" employment or DI/SSI participation status with reports of a severe work limitation. Finally, the claim rejection probit confirms the unconditional results from Table 1: women are more likely to have their disability insurance claim rejected. The parameter λ shows that the variance of the error is lower for women, although the estimate is imprecise.[28]

Disability Vignettes

Table 8 reports summary statistics for the vignette data, separately for men and for women. We present results for all participants who are in the HRS 2007 wave and respond to the vignette questions. Each respondent is asked to rank a vignette in terms of severity of the disability ("Not disabled", "Mildly disabled", "Moderately disabled", "Severely disabled" and "Extremely disabled"). To match the analysis from the first part of the paper, we construct a binary disability indicator, and assume that a vignette is disabled if the respondent answers that the vignette's disability is "Severe" or "Extreme" (we use the same conversion for the self-reports of disability). Table 8 shows that men tend to be more "lenient" than women: they are more likely to classify a vignette as disabled, independently of the gender of the person in the vignette. Further, the gender of the person in the vignette matters: when the vignette is a woman, it is less likely to be classified as disabled. Finally, male respondents are more likely to report being disabled.

[28]The parameter λ captures the shift in the variance of the error term by gender in the heteroskedastic probit model. See Appendix C.

Table 8: Vignettes: Descriptive Statistics

	All resp.		*Male resp.*		*Female resp.*	
	Mean	SD	Mean	SD	Mean	SD
Vignette disabled						
overall	0.38	0.49	0.40	0.49	0.37	0.48
female hypoth. person	0.37	0.48	0.39	0.49	0.36	0.48
male hypoth. person	0.39	0.49	0.40	0.49	0.38	0.48
Respondent disabled	0.11	0.31	0.13	0.33	0.10	0.30
Respondent's age	65.41	10.54	66.31	10.25	64.83	10.69
Observations	40113		15822		24291	

Note: Vignette is classified as disabled if the respondent reports that the vignette is "Severely limited" or "Extremely limited". The same categorization is used for the self-report of disability.

Table 9 presents the results of running probit regressions to estimate equation (16) above. The dependent variable is whether the respondent classifies a given vignette as disabled. In column (1) we control only for a female respondent dummy and dummies for the vignette domain ("Depression", "Pain", or "Cardiovascular disease"); in column (2) we add the female vignette dummy; in column (3) we add additional demographic controls. The female respondent dummy identifies the "pain threshold" parameter ($\gamma_{\bar{L}}$). We find that women respondents are less likely to report that a given vignette is disabled, implying that they have *higher* pain thresholds than men respondents. In columns (2)-(3), we find that respondents tend to be "tougher" on female vignettes: they are less likely to classify a vignette as disabled if that vignette is named "Tamara" as opposed to "Mark", for example. We find no evidence that women respondents tend to be tougher on women vignettes (results not reported).

Table 9: Probit regressions for the probability that vignette is disabled

	(1)	(2)	(3)
Female Respondent	-0.031***	-0.031***	-0.031***
	(0.007)	(0.007)	(0.007)
Female Vignette		-0.017***	-0.016***
		(0.006)	(0.006)
Age			-0.001
			(0.004)
Age squared			0.000
			(0.000)
College degree			-0.004
			(0.007)
Black			0.059***
			(0.010)
Vign. health cond. FE	Yes	Yes	Yes
Rating of own health cond. FE	No	No	Yes
Observations	40113	40113	40077

Note: Standard errors in parentheses, clustered at the individual level. The reported coefficients are marginal effects. ***, **, and * means significance at 1, 5 and 10 percent level, respectively.

Mechanism Parameters

Table 10 reports estimates of the mechanism parameters using optimal minimum distance (see Appendix C): α_L (measuring gender differences in the underlying true, latent work limitations), $\delta_{\bar{A}}$ (gender differences in the cost of applying), $\gamma_{\bar{L}}$ (gender differences in disability perceptions or pain thresholds), θ_{SSA} (which measures differences in disability standards set by SSA for women vs. men), and $\Delta = \sigma^2_\zeta(F_i = 1) - \sigma^2_\zeta(F_i = 0)$ (measuring the marginal effect of gender on the variance of noise in the signal received by SSA). We use the identification scheme described above, and calculate standard errors using the Block Bootstrap (based on 200 replications).

Table 10 shows that estimates for α_L, $\delta_{\bar{A}}$ and $\gamma_{\bar{L}}$ are all positive (although only $\gamma_{\bar{L}}$ is statistically significant).[29] This implies that we can dismiss that lower pain thresholds,

[29]While our estimates suggesting that women have higher pain thresholds are based on self-reports, there is also a vast medical literature that attempts to examine the relationship between pain tolerance/sensitivity and gender using experimental data. In a review of this literature, Racine et al. (2012) conclude: "10 years of laboratory research have not been successful in producing a clear and consistent pattern of sex differences in human pain sensitivity, even with the use of deep, tonic, long-lasting stimuli, which are known to better mimic clinical pain". In a review of both clinical and experimental studies, Fillingim et al. (2009) conclude

Table 10: Structural estimates

Description	Parameter	Estimate
Health distr. shifter	α_L	0.034
		(0.038)
Application threshold	$\delta_{\bar{A}}$	0.019
		(0.052)
Disability report threshold	$\gamma_{\bar{L}}$	0.086***
		(0.020)
SSA threshold	θ_{SSA}	-0.047***
		(0.015)
Signal noise shifter	Δ	-0.842***
		(0.237)
OID test statistics		0.264
		[p-value 61%]

Note: Optimal minimum distance standard errors in parenthesis.

lower application thresholds or less severe impairments for women are the explanation for higher Type I errors. Similarly, the noise of the signal is estimated to be lower for women ($\Delta < 0$), which is inconsistent with the idea that higher Type I errors for women reflect harder-to-verify work limitations.

On the other hand, we estimate a negative value for θ_{SSA}, suggesting a form of statistical discrimination against female applicants. Recall that identification of this parameter comes from two sources: HRS respondents setting higher standards for disability classification when the vignette is a woman, as well as actual rejection rates of disability insurance applicants. This means that the model is overidentified. The test of the single overidentifying restriction shows no evidence of mispecification. The results of Tables 4 and 5 suggest that this form of statistical discrimination arises from having assessed the presence of residual functional capacity rather than the absence of a medical condition.

that "recent clinical and epidemiologic findings generally indicates that women are at increased risk for many chronic pain conditions, and women tend to report higher levels of acute procedural pain" (which of course would make our findings even more puzzling), while "findings regarding sex differences in experimental pain indicate greater pain sensitivity among females compared with males for most pain modalities [...]. The evidence regarding sex differences in laboratory measures of endogenous pain modulation is mixed, as are findings from studies using functional brain imaging to ascertain sex differences in pain-related cerebral activation".

7 Conclusions

This paper documents substantial differences in Type I error across genders. In particular, we find that women with a severe, work-related, permanent impairment are more likely to have their disability insurance application turned down (i.e., suffer a Type I error) than men with observationally equivalent characteristics. We show that supply considerations (such as how the SSA screens applicants) are more plausible explanations than demand-side channels (such as differences in disability perceptions or application costs).

These false rejections have lasting consequences. The higher false rejections for women are arising because of an assessment that women have residual work capacity. We show that despite this assessment, these women are not returning to work. By contrast, those women who have been "correctly" rejected return to work at a significantly higher rate. Further, we do not find such a difference among men who have been rejected, implying that their residual work capacity was more accurately assessed.

One of the surprising aspects of studying DI/SSI application forms is the amount of information on those forms which may not be relevant to assessing the extent of a work limitation. For example, the form asks not only about gender, but also about marital status. Our results suggest that an important policy change to consider would be to make disability insurance applications gender-blind.[30] Evidence from other settings show that gender-blind evaluations of candidates matter for explaining a variety of labor market outcomes (Rouse and Goldin, 2000; Bertrand and Mullainathan, 2004; Card et al., 2019). However, there are two caveats to making the evaluation for disability insurance completely gender blind: first, some illnesses are readily associated with gender; second, the same illness may create different degrees of incapacity by gender, as recent evidence on gender-based medicine suggests. Nonetheless, our finding of substantial gender-based errors raises important policy issues.

[30] Alternatively, to consider machine learning algorithms that take out all human biases.

References

Autor, D., A. R. Kostøl, M. Mogstad, and B. Setzler (2019, July). Disability Benefits, Consumption Insurance, and Household Labor Supply. *American Economic Review 109*(7), 2613–2654.

Autor, D. H., N. Maestas, K. J. Mullen, and A. Strand (2015). Does Delay Cause Decay? The Effect of Administrative Decision Time on the Labor Force Participation and Earnings of Disability Applicants. Working Paper 20840, National Bureau of Economic Research.

Bangasser, D. A., S. R. Eck, and E. O. Sanchez (2019). Sex differences in stress reactivity in arousal and attention systems. *Neuropsychopharmacology 44*(1), 129–139.

Benitez-Silva, H., M. Buchinsky, H.-M. Chan, J. Rust, and S. Sheidvasser (2004). How Large is the Bias in Self-Reported Disability Status? *Journal of Applied Econometrics 19*(6), 649–670.

Benitez-Silva, H., M. Buchinsky, and J. Rust (2004). How Large are the Classification Errors in the Social Security Disability Award Process? NBER Working Papers 10219, National Bureau of Economic Research.

Bertrand, M. and S. Mullainathan (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review 94*(4), 991–1013.

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

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2019). Are Referees and Editors in Economics Gender Neutral? *Quarterly Journal of Economics*. Forthcoming.

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

Clocchiatti, A., E. Cora, Y. Zhang, and P. Dotto (2016). Sexual dimorphism in cancer. *Nature Reviews Cancer 16*(5), 330–339.

Collins, K. P. and A. Herfle (1985). Social Security Disability Reform Act of 1984: Legislative History and Summary of Provisions. *Social Security Bulletin 48*(4), 5–11.

Daly, M. and R. V. Burkhauser (2003). The Supplemental Security Income Program. In R. Moffitt (Ed.), *Means-Tested Transfer Programs in the United States*, NBER Chapters, pp. 79–140. National Bureau of Economic Research, Inc.

Duggan, M. and S. A. Imberman (2009). Why are the Disability Rolls Skyrocketing?

The Contribution of Population Characteristics, Economic Conditions, and Program Generosity. In D. Cutler and D. Wise (Eds.), *Health at Older Ages: The Causes and Consequences of Declining Disability among the Elderly*, Chapter 11, pp. 337–379. University of Chicago Press.

Fillingim, R., C. King, M. Ribeiro-Dasilva, B. Rahim-Williams, and J. R. III (2009). Sex, gender, and pain: A review of recent clinical and experimental findings. *The Journal of Pain: Official Journal of the American Pain Society 10*(5), 447–485.

Hancock, M.-A. (2004). *The Politics of Disgust: The Public Identity of the Welfare Queen*. New York University Press.

Haveman, R. and B. Wolfe (2000). The economics of disability and disability policy. In A. J. Culyer and J. P. Newhouse (Eds.), *Handbook of Health Economics* (1 ed.), Volume 1, Chapter 18, pp. 995–1051. Elsevier.

Kapteyn, A., J. P. Smith, and A. van Soest (2007). Vignettes and Self-Reports of Work Disability in the United States and the Netherlands. *American Economic Review 97*(1), 461–473.

Kostøl, A. R. and M. Mogstad (2014). How Financial Incentives Induce Disability Insurance Recipients to Return to Work. *American Economic Review 104*(2), 624–655.

Legato, M. J., P. A. Johnson, and J. Manson (2016). Consideration of Sex Differences in Medicine to Improve Health Care and Patient Outcomes. *Journal of the American Medical Association 316*(18), 1865–1866.

Low, H. and L. Pistaferri (2015). Disability Insurance and the Dynamics of the Incentive Insurance Trade-Off. *American Economic Review 105*, 2986–3029.

Low, H. and L. Pistaferri (2019). Disability Insurance: Theoretical Trade-Offs and Empirical Evidence. Technical report.

Michaud, A. and D. Wiczer (2018). Occupational hazards and social disability insurance. *Journal of Monetary Economics 96*, 77–92.

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

Neumark, D., R. J. Bank, and K. D. V. Nort (1996). Sex Discrimination in Restaurant Hiring: An Audit Study. *The Quarterly Journal of Economics 111*(3), 915–941.

Racine, M., Y. Tousignant-Laflamme, L. A. Kloda, D. Dion, G. Dupuis, , and M. Choinière (2012). A systematic literature review of 10 years of research on sex/gender and experimental pain perception – Part 1: Are there really differences between women and men? *Pain 153*, 602–618.

Rouse, C. and C. Goldin (2000). Orchestrating Impartiality: The Impact of Blind Auditions on Female Musicians. *American Economic Review 90*(4), 715–741.

Sarsons, H. (2019). Gender Differences in Recognition for Group Work. *Journal of Political Economy*. Forthcoming.

United States General Accounting Office (1994). Social Security Disability. Most of Gender Difference Explained. Report to the Ranking Minority Member, Special Committee on Aging, U.S. Senate GAO/HEHS-94-94.

A Appendix: Additional Figures and Tables

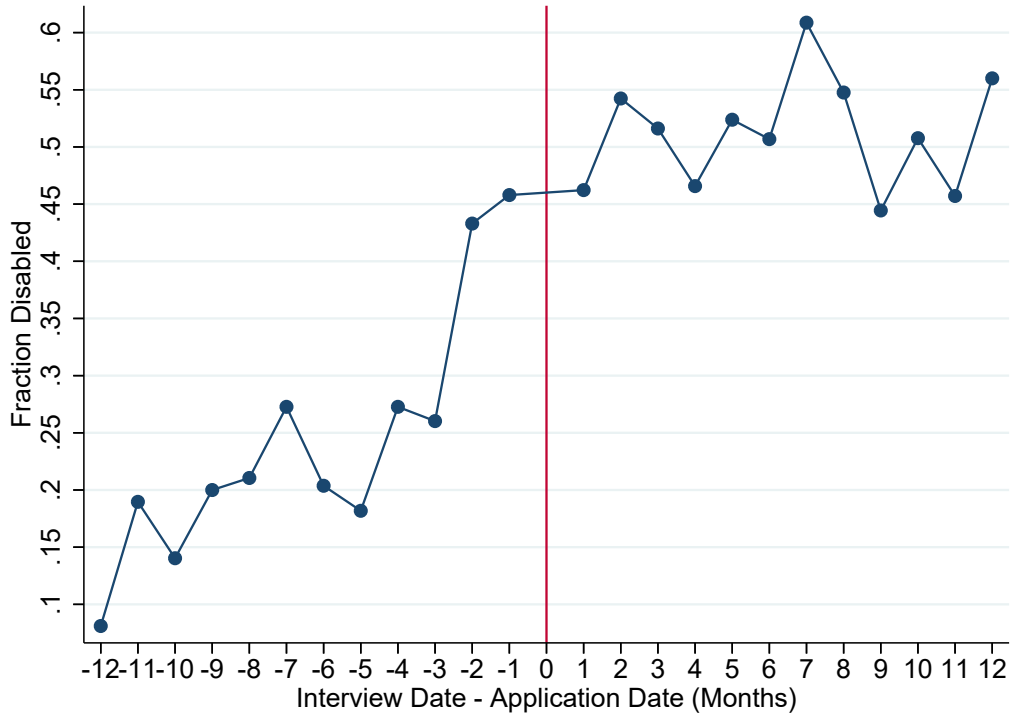

Figure A.1: Fraction reporting a work limitation by distance between HRS interview and disability insurance application date

Table A.1: Additional results

	(1)	(2)	(3)	(4)
Female	0.213***	0.185***	0.313***	0.201***
	(0.051)	(0.054)	(0.078)	(0.052)
Age 50-55	-0.074			
	(0.094)			
Age 55-59	-0.363***			
	(0.083)			
Age 60-65	-0.287***			
	(0.087)			
Phys.occ.req. index		-0.004		
		(0.008)		
Female*empl. husb.			-0.046	
			(0.063)	
Female*married			-0.130	
			(0.105)	
Other demographics	Yes	Yes	Yes	Yes
Health cond. FE	Yes	Yes	Yes	Yes
HRS Objective FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
ADL FE	Yes	Yes	Yes	Yes
BMI+Hosp	Yes	Yes	Yes	Yes
Occupation FE	Yes	No	Yes	Yes
Observations	447	415	447	447

Note: Standard errors in parentheses, clustered at the individual level. Coefficients are estimates of marginal effects. Other demographics include College degree, Black, Years of labor market experience, SSI applicant, Concurrent SSI/DI applicant, Married, Widowed, and Age (except in column (1)). Column (1) replaces age with an age spline. Column (2) replaces occupation dummies with a physical occupational requirement index using a mapping between HRS occupational codes and O*NET data (as in Michaud and Wiczer, 2018). Column (3) adds the interaction of the female dummy with married female and a female with an employed husband. Column (4) is Type I error regression including DDS reconsideration.

Table A.2: Disability transition, ages 25-65

	(1)	(2)	(3)
Disabled at $t-1$	0.5217***	0.5180***	0.3541***
	(0.0078)	(0.0133)	(0.0144)
Female		0.0105***	0.0053**
		(0.00120)	(0.0022)
Disabled at $t-1 \times$ female		0.0053	0.0205
		(0.0164)	(0.0171)
College degree			-0.0112***
			(0.0023)
Black			0.0121***
			(0.0034)
Married			0.0050*
			(0.0027)
Widowed			0.0144***
			(0.0052)
Age			0.0006***
			(0.0002)
Health conditions FE	No	No	Yes
Year FE	No	No	Yes
ADL FE	No	No	Yes
BMI+Hosp. FE	No	No	Yes
Occupation FE	No	No	Yes
Observations	60831	60831	52195

Note: Dependent variable is whether the individual is disabled at time t. Standard errors in parentheses, clustered at the individual level.

B Appendix: The Effect of Gender on Type I Errors

Section 3 outlines a framework for the mechanisms through which gender might lead to differences in Type I errors: differences in underlying health, in pain thresholds, in application thresholds, differences in the SSA assessment of health and differences in signal precision.

These mechanisms are summarized by the following equations:

$$L_i^* = \alpha_0 + \alpha_L F_i + \varepsilon_i \tag{B.1}$$

$$\bar{L}_i = \gamma_0 + \gamma_{\bar{L}} F_i \tag{B.2}$$

$$\bar{A}_i = \bar{L} + \delta_0 + \delta_{\bar{A}} F_i \tag{B.3}$$

$$S_i^* = L_i^* + \theta_{SSA} F_i + \zeta_i \tag{B.4}$$

where the signal that the SSA observes, S_i^*, is assumed to be a noisy realization of the true health status, L_i^*. The signal S_i^* can be rewritten using the equation for L_i^* as:

$$S_i^* = \alpha_0 + (\alpha_L + \theta_{SSA}) F_i + \varepsilon_i + \zeta_i \tag{B.5}$$

The SSA decision is:

$$\textit{Award if: } S_i^* > \bar{L}_{SSA} \tag{B.6}$$

We use these equations to characterize Type I errors. We then simulate how Type I errors vary with each of the key parameters: $\left\{\alpha_L, \gamma_{\bar{L}}, \delta_{\bar{A}}, \theta_{SSA}, \sigma_\zeta^2(F)\right\}$, respectively representing gender differences in underlying health, pain thresholds, application thresholds, differences in the SSA assessment of an applicant's health, and gender-dependent noise variance in work limitations signal.

Type I Errors

In principle, we can distinguish the case where everyone who has $L^* > \bar{L}$ applies from the case where the application threshold lies to the right of $\bar{L}$, and so everyone who applies has $L^* > \bar{L}$, but not everyone with $L^* > \bar{L}$ applies (due to, say, opportunity costs from applying). However, the linearity in equations (B.2) and (B.3) means that we can subsume the effect on Type I errors of $\delta_{\bar{A}}$ into the effect of $\gamma_{\bar{L}}$.

We define the Type I error as:

$$\begin{aligned}\Pr\left(reject|disabled\ appl.\right) &= \frac{\Pr\left(S^* < \bar{L}_{SSA}\ \&\ L^* > \bar{L}\right)}{\Pr\left(L^* > \bar{L}\right)} \\ &= \pi(.)\end{aligned}$$

This can be rewritten as:

$$\pi(.) = \frac{\Pr\left(\alpha_0 + (\alpha_L + \theta_{SSA})\,F + \varepsilon + \zeta < \bar{L}_{SSA},\ \ \alpha_0 + \alpha_L F + \varepsilon > \gamma_0 + \gamma_{\bar{L}} F\right)}{\Pr\left(\alpha_0 + \alpha_L F + \varepsilon > \gamma_0 + \gamma_{\bar{L}} F\right)}$$

$$= \frac{\Pr\left(\begin{array}{c} -\infty < \varepsilon + \zeta < \left(\bar{L}_{SSA} - \alpha_0\right) - (\alpha_L + \theta_{SSA})\,F, \\ (\gamma_0 - \alpha_0) + (\gamma_{\bar{L}} - \alpha_L)\,F < \varepsilon < \infty \end{array}\right)}{1 - \Pr\left(\varepsilon < (\gamma_0 - \alpha_0) + (\gamma_{\bar{L}} - \alpha_L)F\right)}$$

Finally, we assume joint normality of the error terms but consider the case in which the variance of the noise ζ can be gender-dependent:

$$\begin{pmatrix} \varepsilon \\ \zeta \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & \sigma_\zeta^2(F) \end{pmatrix}\right),$$

The expression for the Type I error can thus be rewritten as:

$$\pi(.) = \frac{\Phi\left(\frac{(\bar{L}_{SSA} - \alpha_0) - (\alpha_L + \theta_{SSA})F}{\sqrt{\sigma_\varepsilon^2 + \sigma_\zeta^2(F)}}\right) - G\left(\frac{(\bar{L}_{SSA} - \alpha_0) - (\alpha_L + \theta_{SSA})F}{\sqrt{\sigma_\varepsilon^2 + \sigma_\zeta^2(F)}}, \frac{(\gamma_0 - \alpha_0) + (\gamma_{\bar{L}} - \alpha_L)F}{\sigma_\varepsilon}; \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + \sigma_\zeta^2(F)}}\right)}{1 - \Phi\left(\frac{(\gamma_0 - \alpha_0) + (\gamma_{\bar{L}} - \alpha_L)F}{\sigma_\varepsilon}\right)} \tag{B.7}$$

where $\Phi(.)$ and $G(.,.;\rho)$ are the CDF's of the standard and jointly standard normal distribution with correlation coefficient ρ, respectively.

We cannot sign the derivatives of equation (B.7) for the general case. In particular, the distribution of errors will matter for the way each mechanism impacts on Type I errors. Instead, we illustrate the channels using simulation.

Define $\Delta = \sigma_\zeta^2(F = 1) - \sigma_\zeta^2(F = 0)$ as the "marginal effect" of the female dummy on the variance of the signal observed by the SSA. A positive value of Δ means that signal for women is more noisy than for men. Figure B.2 shows how Type I errors differ for different values of $\{\alpha_L, \gamma_{\bar{L}}, \theta_{SSA}, \Delta\}$. The value where each line cuts the y-axis is the Type I error for

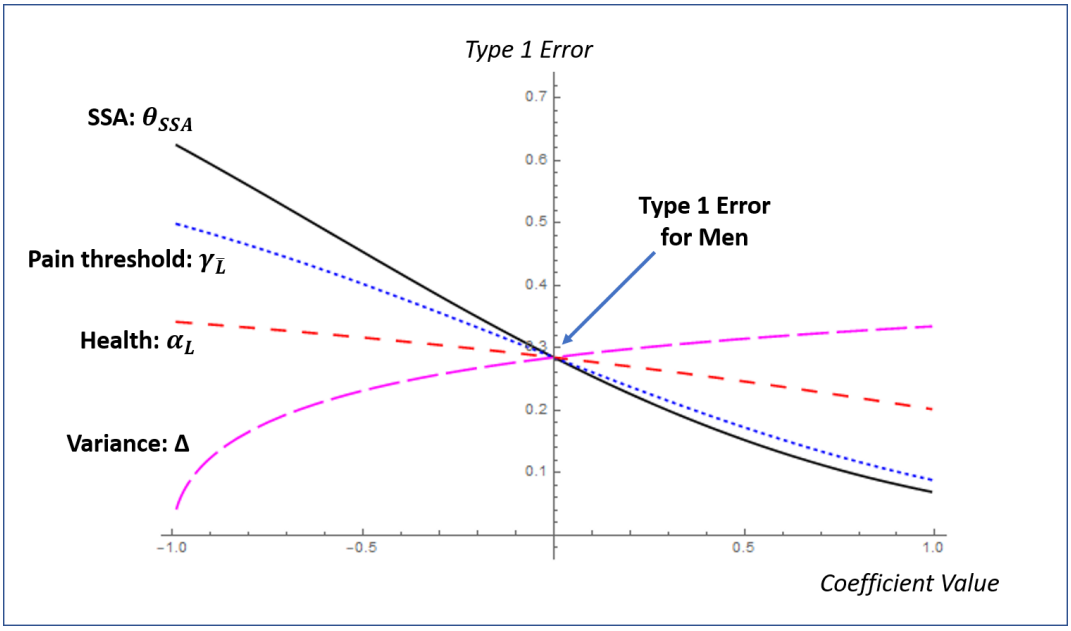

Figure B.2: Simulated Type I Errors

men.[31] Negative values for each of the coefficients $\{\alpha_L, \gamma_{\bar{L}}, \theta_{SSA}\}$ indicate, respectively, that women do not have as severe an underlying work limitation, that they have a lower pain threshold, or that the SSA assesses their work limitation as being less severe than men. The figure shows that negative values for any of these parameters leads to greater Type I errors for women than for men. Our estimates in section 6 show that only θ_{SSA} is (statistically significantly) negative. A negative value for Δ means that the SSA has a more precise signal of women's health than it has of men's health. This more precise signal leads to a lower Type I error for women than men. Our estimate of Δ in section 6 is negative and so this is not the explanation of the higher Type I error rate.

[31] We experiment with different values of α_0, γ_0 and $\bar{L}_{SSA}$. The qualitative results remain, although the level of Type I errors varies.

C Appendix: Minimum Distance Estimation

We identify the structural parameters of the model using a simple Minimum Distance procedure. Recall that our structural equations (see Section 6) are:

$$\begin{aligned}
\Pr(L_i = 1|X_i, F_i) &= \Phi((\alpha_0 - \gamma_0) + X_i'\alpha_x + (\alpha_L - \gamma_{\bar{L}})F_i) \\
\Pr(A_i = 1|X_i, F_i) &= \Phi((\alpha_0 - \gamma_0 - \delta_0) + X_i'(\alpha_x - \delta_x) + (\alpha_L - \gamma_{\bar{L}} - \delta_{\bar{A}})F_i) \\
\Pr(R_i = 1|X_i, F_i) &= \Phi((1 + \sigma_\zeta^2(F_i))^{-1/2}((\bar{L}_{SSA} - \alpha_0) - X_i'\alpha_x - (\alpha_L + \theta_{SSA})F_i) \\
\Pr(L_{v,i} = 1|X_i, F_v, F_i) &= \Phi(\theta_v - \gamma_0 + X_i'\theta_x + \theta_{SSA}F_v - \gamma_{\bar{L}}F_i)
\end{aligned}$$

The corresponding reduced form equations are:

$$\begin{aligned}
\Pr(L_i = 1|X_i, F_i) &= \Phi(b_0^L + X_i'b_1^L + b_2^L F_i) \\
\Pr(A_i = 1|X_i, F_i) &= \Phi(b_0^A + X_i'b_1^A + b_2^A F_i) \\
\Pr(R_i = 1|X_i, F_i) &= \Phi(e^{-\lambda F_i}(b_0^R + X_i'b_1^R + b_2^R F_i)) \\
\Pr(L_{v,i} = 1|X_i, F_v, F_i) &= \Phi(b_0^V + X_i'b_1^V + b_2^V F_v + b_3^V F_i)
\end{aligned}$$

This establishes the following mapping between structural and reduced form parameters:

$$\begin{pmatrix} \alpha_L - \gamma_{\bar{L}} \\ \alpha_L - \gamma_{\bar{L}} - \delta_{\bar{A}} \\ -(\alpha_L + \theta_{SSA}) \\ \theta_{SSA} \\ -\gamma_{\bar{L}} \\ \frac{\ln(1+\Delta)}{2} \end{pmatrix} = \begin{pmatrix} b_2^L \\ b_2^A \\ b_2^R \\ b_2^V \\ b_3^V \\ \lambda \end{pmatrix} \tag{C.1}$$

where $\Delta = \sigma_\zeta^2(F = 1) - \sigma_\zeta^2(F = 0)$ is the "marginal effect" of the female dummy on the variance of the noise.

Call $\hat{b}(\beta)$ the reduced form estimates. We obtain the structural estimates β by solving:

$$\min_\beta (\hat{b}(\beta) - \beta)'\Omega(\hat{b}(\beta) - \beta)$$

where Ω is the optimal weighting matrix, corresponding to the inverse of the variance matrix of $\hat{b}(\beta)$, which we obtain by the block bootstrap (with 200 replications).