NBER WORKING PAPER SERIES

DISABILITY INSURANCE: ERROR RATES AND GENDER DIFFERENCES

Hamish Low
Luigi Pistaferri

Working Paper 26513
http://www.nber.org/papers/w26513

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2019, Revised May 2024

Thanks to Sarah Eichmeyer, Max Rong, and Charlotte Woodacre for invaluable research assistance, to five anonymous referees, David Autor, David Card, Stephen Haider, Hanming Fang, Amanda Michaud, Magne Mogstad, John Rust, Lucie Schmidt, Alessandra Voena, and participants at various seminars and conferences for comments. This paper uses restricted HRS data made available to Pistaferri under confidential agreements RDA 2015-028 and RDA 2020-070. All errors are ours. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Disability Insurance: Error Rates and Gender Differences
Hamish Low and Luigi Pistaferri
NBER Working Paper No. 26513
November 2019, Revised May 2024
JEL No. H55,J16,J71

ABSTRACT

We show the extent of screening errors made in disability insurance awards using matched survey-administrative data. Type I errors are widespread with large gender differences. Work-disabled women are 12.8 percentage points more likely to be rejected than work-disabled men, controlling for health conditions and demographics. Gender differences arise because women are assessed with more residual work capacity. We model the SSA decision-making process and estimate that gender differences in screening errors originate from lower utility losses from incorrectly rejecting women. Finally, noise in self-reported work limitation leads to an overstatement of screening errors, but the gender difference remains.

Hamish Low
Department of Economics
University of Oxford
Oxford England
hamish.low@economics.ox.ac.uk

Luigi Pistaferri
Department of Economics
579 Jane Stanford Way
Stanford University
Stanford, CA 94305-6072
and NBER
pista@stanford.edu

1 Introduction

Disability insurance is a major part of social insurance provision in the US and elsewhere. There is substantial evidence on the labor supply disincentive effects of the program, but there is relatively less 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 measure the difference in false rejections between men and women and to explain why it arises.

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 managed 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. Reflecting these concerns, there exists a sizable literature that has looked at the labor supply incentive consequences of disability insurance.[2] 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 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

[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 fast increase in caseloads. The programs differ in 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 2019 the DI program paid cash benefits of around $145 billion; in the same year, total SSI expenditure was $56 billion, absorbing 16% of federal non-Medicaid welfare spending. In terms of growth of recipiency, between 1984 and 2019 the share of disabled workers receiving DI benefits out of all workers increased from 2.4% to 5.4%. As for SSI, the fraction of 18-64 years old who receive SSI benefits has doubled from 1.2% in 1984 to 2.3% in 2019.

[2]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 (2020).

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).[3]

Other inefficiencies may come from the fact that disability assessment is, at least in part, subjective. This opens up the possibility of bias – even if unintended – against specific demographic groups such as women. Alternatively, as discussed in the medical literature (Legato et al., 2016; Bangasser et al., 2019) 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.[4] 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. Cabral and Dillender (2021) find that female patients randomly assigned a female doctor rather than a male doctor are more likely to be evaluated as disabled when applying for workers' compensation and to be awarded subsequent cash benefits. 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.

Estimating screening errors requires a measure of "true" health-related work limitations alongside "reported to the authorities" work disabilities. We start by merging information from the Health and Retirement Study (HRS) with administrative data from the SSA on DI and SSI applications and social security earnings.[5] To our knowledge, we are the first to use these linked data to study the efficiency aspects of the DI/SSI programs. Administrative data

[3]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.

[4]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 2019 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).

[5]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). The merging of application data from the two programs makes inference more reliable because in survey data the number of applicants to either program is typically small.

allow us to observe the application process (its filing, dates, outcomes, and justifications) without error. Survey data provide information on self-reported work disability that match as close as possible the institutional definition: a severe, non-temporary medical disability that prevents working. Below, Type I errors correspond to applicants who self-report to be severely work limited being turned down; Type II errors correspond to applicants who self-report not to be severely work limited being awarded benefits.

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. Our study is the first to document this finding.[6] This main result 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 having residual capacity for work.

We propose a model of the decision-making process by the SSA that leads to Type I errors. The SSA makes award decisions using a noisy indicator of work limitation to minimise an objective defined over costly screening errors. Differences in Type I errors by gender may arise through three channels: differences in demand for disability insurance by gender; differences in the cost of screening errors by gender; or differences in the distribution of observables between men and women. Demand may differ between men and women due to gender differences in the severity of actual work-related impairments or perceived disability norms, or differences in terms of opportunity costs of applying. Further, these differences in demand will feed into the SSA decision and generate statistical discrimination. On the supply side, women may have observable characteristics that induce higher denial rates due to existing program rules; the SSA may receive a noisier signal from female applicants; or have a lower disutility from false rejections of work-limited women and a higher disutility of false acceptances of non-work limited women, which we interpret as taste-based discrimination. Our definition of taste-based discrimination does not mean chauvinism (or "animus" towards women). It simply means that the utility losses from type I errors among women are (in relative terms) lower than those from type I errors among men - but both utility losses could

[6] A study by the United States General Accounting Office (1994) reported higher DI 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.

be substantial. An alternative interpretation of our findings is that SSA evaluators have the distorted belief that women are more likely to have residual functional capacity than men, conditional on other observables, and this feeds into higher denial rates (even without taste bias).

To assess these explanations, we use data on self-reports of work disability, application rates, rejection rates, and out-of-pocket health spending. 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 respondents assess the relation between gender and work limitations in the general population.

We use our estimates to decompose the sources of the gender differences in Type I errors. Differences may arise because of differences between men and women in other observables, such as occupation or health status. Indeed, some of these differences are written into the process through legislation. We distinguish the impact of these institutional differences from taste-based and statistical discrimination. Our key conclusion is that differences in the award thresholds for men and women due to taste-based discrimination explain a substantially larger fraction of the observed gender difference than differences in the distribution of observables or statistical discrimination.

An alternative way of testing for taste-based discrimination is to use an "outcome test" strategy (Becker, 1971). If admission standards are higher for women due to lower utility losses from rejecting them, incorrectly rejected women should be on average in worse health - and hence less likely to work - than incorrectly rejected men. We find that post-decision employment rates of incorrectly rejected women are lower than those of incorrectly rejected men, confirming the main finding of the structural analysis. This result does not reflect heterogeneity: in the years preceding the disability shock the two groups have similar employment rates.

There are only a few papers that estimate classification errors associated with disability insurance. Nagi (1969) uses a sample of 2,454 DI applicants that were assessed by a team of medical professionals independent of the SSA and compares such assessment to the SSA decision. 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 main 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. Since these are hard-to-verify conditions, classification errors in the post-1984 era may be very different.

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 an 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 (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 HRS waves, disability application questions are only asked to those reporting a disability, which induces a mechanical understatement of Type II errors. Finally, even some truly disabled applicants may be rejected because of not meeting eligibility requirements (e.g., not having contributed for enough years to the system). In survey data this may be incorrectly classified as Type I error even though it reflects program rules. We improve over previous literature on several aspects. First, we use multiple questions from the HRS to get at a definition of work disability that mirrors as closely as possible the institutional definition of work disability used by SSA; second, we allow self-reports to measure the true work disability with error and show how

much this biases estimates of screening error; third, we use administrative data to define eligibility.

The rest of the paper proceeds as follows. In Section 2 we provide institutional details on the programs that insure against work limitation shocks, present the data, and explains how we measure work limitations. Section 3 discusses our estimates of screening errors and various extensions. Sections 4, 5 and 6 present a simple theoretical framework for explaining the empirical findings on error rates, discuss identification and results for the structural parameters of the model, and discusses various implications of Type I errors, respectively. Section 7 concludes.

2 Institutional Details and Data

2.1 The DI and SSI programs

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.[7] The difficulty with providing insurance is 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.[8] 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 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

[7]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.

[8]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).

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, $1,220 a month for non-blind individuals as of 2019); and (5) the individual must meet a medical requirement preventing work. Requirement (2) suggests that the DI program is designed by law to insure individuals who have worked in the market and not those who have worked from home. This implicitly generates less disability insurance for women than men, but this is not the type of gender differences we are concerned about here. Indeed, as we explain below, applicants for DI benefits in our administrative data only include those who satisfy requirement (2) (those who don't are issued "technical denials" and do not appear in the administrative records). Requirement (5) is the same as in the SSI program, and we discuss it below after a short description of SSI.

Working-age individuals who are deemed to be disabled and have limited income and limited resources are eligible to receive supplemental security income (SSI).[9] 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.[10] SSI benefits are adjusted annually. In 2019, an individual (couple) with no countable income would receive $771 ($1,157) in cash benefits a month.

2.2 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, or Disability Determination Service (DDS). The case is assigned in a quasi-random fashion to an adjudicative team consisting of a medical or psychological consultant and a disability examiner (Maestas et al., 2013). 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 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

[9]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.

[10]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.).

impairment" the individual is awarded benefits without further review.[11] 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 35% of DI/SSI applicants are awarded benefits at the initial DDS stage.[12] But rejections can be appealed and about 56% 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 (11%). Those denied at the reconsideration stage can further appeal (and 80% do so). These appeals are decided outside of the DDS, by Administrative Law Judges (ALJ) or at higher hearing levels, and applications tend to have a much larger success rate (68%).[13]

2.3 Data: The Health and Retirement Study (HRS)

The Health and Retirement Study (HRS) is a panel data set sponsored by the National Institute of Aging (grant n. U01AG009740) and conducted by 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 2020, 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 (see below for the precise definition); (b) indicators for the presence of specific health conditions (high blood pressure, diabetes,

[11] 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).

[12] Average data from 1992-2012 (source: 2020 Annual Statistical Report on the Social Security Disability Insurance Program, Tables 61-63.)

[13] 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 (see Hoynes et al., 2022).

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; (c) out-of-pocket individual medical spending information (unavailable in the first two HRS waves); and (d) Disability Vignette data (available in a special module of the 2007 wave).

2.4 Data: Social Security Administration Records

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 include "technical denials" (e.g., applications denied for non-medical reasons, such as applicants not having accumulated enough work credits to be insured for SSDI) or information on decisions made at the ALJ level and beyond.[14]

The F831 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). Individuals can have several application cycles and at most two rounds per cycle. 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;[15] (d) the stage at which the application is denied (or awarded); (e) whether it is a DI, an SSI, or a concurrent DI/SSI application; and (f) other miscellaneous information, such as whether the applicant was asked to submit to a

[14]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: "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.

[15]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.

consultative examination (which occurs if the applicant's own medical sources are inadequate to determine if he or she is disabled).

2.5 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 approximate 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?". Unless otherwise noted, we classify as work disabled (or severely work limited) people who answer "Yes" to the first and third question and 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. We discuss below alternative definitions of work disability; when we introduce the formal model, we allow self-reports to measure the true work disability with error. Appendix A details how these various disability indicators are defined, while Appendix B discusses the validity of the self-reported disability measures (see in particular Table B.1).

To get a gauge of the validity of work limitation self-reports as capturing (albeit with some noise) true screening errors, in Table B.2 we test whether applicants who have suffered a Type I error (i.e., rejected applicants who self-report a work disability) work as much as correctly rejected applicants. This would be consistent with the SSA assessment that *all* rejected applicants have residual functional capacity to work. Column (1) reports the results of this test, using as dependent variable an indicator for whether the individual had any employment in the three years following the initial consideration stage (results are similar if we look at five year post-decision outcomes).[16] Our main finding is that rejected applicants who self-report a work limitation work significantly less than those who do not.

[16]We define employment in a given year as the individual having 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. The test is an extension of the strategy of Bound (1989) and von Wachter et al. (2011), who compare the labor market outcomes of rejected and non-rejected applicants, by separating further the group of rejected applicants into severely and non-severely work limited applicants.

We interpret this as evidence that Type I errors are true errors: the SSA has overestimated the residual functional capacity of applicants who self-report work limitations.

3 Screening Errors: Estimates and Determinants

We start by providing evidence on the extent of Type I and Type II errors and their determinants. In the next section we present a model with alternative channels that can rationalize this evidence, and interpret the evidence in terms of institutional and legislative differences compared to statistical and taste-based discrimination.

3.1 Sample Statistics

The estimation sample for studying determinants of screening errors consists of HRS (non-proxy) respondents aged 20-65 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. 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.[17]

In Table 1 we report descriptive statistics for the matched sample, comprising 1,605 first-round applications.[18] 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 "justification" story (i.e., people rationalizing their application for disability insurance by over-reporting work limitations, Bound and Burkhauser, 1999) and more likely to be consistent with the idea that people report truthfully their health conditions to HRS

[17] If we include self-reports of work disability that happen much earlier than the application 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 $-12 \leq d < -2$, jumps upward at $d = -2$ (to about 35%), and remains approximately around 50% for $-2 < d \leq 12$ (see Figure A.1 in the Appendix). In Table 4, we report results using the $-2 \leq d \leq 12$ sample and show that they are similar to the baseline.

[18] There are 645 from men and 960 from women. The larger female share 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.

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 work disability (the binary indicator "Some work limitation" based on the question: "Do you have any impairment or health problem that limits the kind or amount of paid work you could do?"), the fraction of applicants to DI/SSI with a work disability jumps to 83%. 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 very closely the denial rate observed in the population of all applicants at the initial consideration stage (61%). The (cumulative) denial rate is slightly lower if we also consider the reconsideration stage (57%). In the raw data, there are large differences in the denial rate between women and men (a 11 percentage point difference). However, this alone does not indicate a gender difference in classification errors.

In the raw data reported in Table 1, Type I errors (estimated as the fraction of work disabled applicants who are turned down) are large (more than half of applicants who report to be work disabled are rejected). Furthermore, there are large differences in Type I errors between men and women (18 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, less likely to be black, unmarried or widowed, more likely to be employed in blue collar than in clerical or service occupations, and more likely to apply for disability insurance because of a cardiovascular condition, while women are slightly more likely to apply because of mental or respiratory disorders. Nonetheless, Figure 1 shows that gender differences in Type I error exist independently of occupation or the type of disability.

3.2 Gender Differences in Classification Errors: Baseline Estimates

We estimate the effect of gender and other characteristics on Type I errors by running the following probit model for applicants who report to be work disabled ($L_i = 1$):

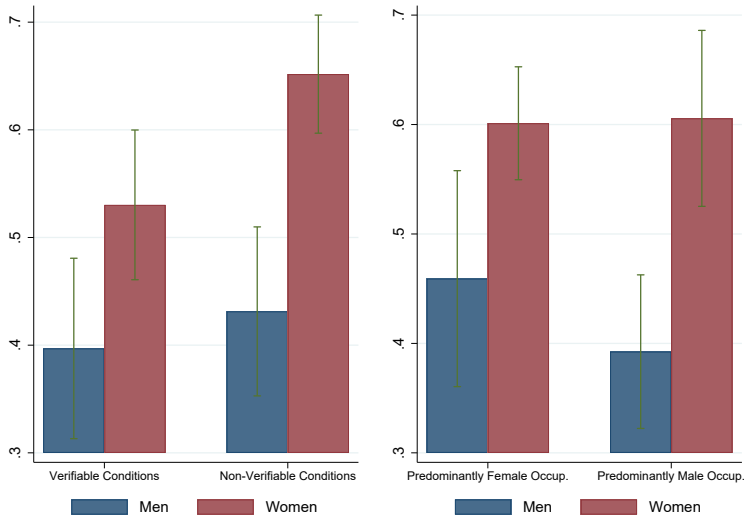
$$\Pr(R_i = 1 | L_i = 1, X_i, F_i) = \Phi(X_i'\psi_0 + \psi_1 F_i)$$

Table 1: Descriptive statistics

	Men		*Women*		*All*	
	Mean	SD	Mean	SD	Mean	SD
Denial, appl. round	0.54	0.50	0.65	0.48	0.61	0.49
Type I error, appl. round	0.42	0.49	0.60	0.49	0.53	0.50
Type II error, appl. round	0.37	0.48	0.30	0.46	0.33	0.47
Denial, appl. cycle	0.50	0.50	0.61	0.49	0.57	0.50
Type I error, appl. cycle	0.35	0.48	0.55	0.50	0.48	0.50
Type II error, appl. cycle	0.39	0.49	0.32	0.47	0.35	0.48
Work disabled	0.44	0.50	0.51	0.50	0.48	0.50
Some work limitation	0.82	0.39	0.84	0.37	0.83	0.38
Applied SSI only	0.24	0.43	0.27	0.45	0.26	0.44
Applied DI + SSI	0.20	0.40	0.22	0.41	0.21	0.41
College degree	0.35	0.48	0.32	0.47	0.33	0.47
Black	0.31	0.46	0.33	0.47	0.32	0.47
Married	0.56	0.50	0.45	0.50	0.49	0.50
Widowed	0.04	0.19	0.11	0.31	0.08	0.27
Lab. mark. experience	32.89	9.90	26.44	10.65	29.04	10.83
Age	57.14	4.39	55.84	5.93	56.36	5.40
Clerical occupation	0.23	0.42	0.39	0.49	0.33	0.47
Services occupation	0.13	0.34	0.31	0.46	0.24	0.43
Blue collar occupation	0.53	0.50	0.18	0.39	0.32	0.47
Type of condition in F831						
Musculoskeletal	0.41	0.49	0.41	0.49	0.41	0.49
Respiratory	0.05	0.21	0.07	0.26	0.06	0.24
Cardiov.	0.17	0.37	0.11	0.31	0.13	0.34
Endocrine	0.05	0.22	0.06	0.24	0.06	0.24
Neurol.	0.08	0.27	0.07	0.26	0.08	0.27
Mental dis.	0.09	0.29	0.10	0.30	0.09	0.29
Cancer	0.03	0.18	0.04	0.20	0.04	0.19
Immune def.	0.03	0.16	0.02	0.15	0.03	0.16
Dig. & Urin.	0.03	0.17	0.03	0.17	0.03	0.17
Other	0.07	0.25	0.08	0.27	0.07	0.26
Number of obs.	645		960		1605	

Note: The sample is HRS respondents that are observed in the F831 dataset in their first-round applications (across all application cycles). An application cycle includes the initial consideration and, if a rejection is appealed, a reconsideration. Respondents are defined as having "some work limitation" if they report to have an impairment or health problem that limits the kind or amount of paid work they can do; and to be "work disabled" if this 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).

Figure 1: Type I error rates by primary disability code and gender



Note: Non-verifiable conditions include "Muscoloskeletal", "Mental", and "Other" disabilities. Predominantly female occupations include clerical and service occupations. The vertical lines are 95% confidence intervals. See Appendix A for variable definitions.

For Type II errors, we look at applicants who do not report a work disability ($L_i = 0$):

$$\Pr(R_i = 0 | L_i = 0, X_i, F_i) = \Phi(X_i'\kappa_0 + \kappa_1 F_i)$$

where i is individual, R_i is a dummy for having a disability insurance application denied, X_i includes individual 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 three columns of Table 2 report results for Type I errors; the last three columns focus on Type II errors. Columns (1)-(3), ordered in terms of richness of controls, show statistically significant higher Type I error rates for women: a 12.8 percentage point difference in the richest specification of column (3). Older applicants are less likely to be turned down if truly disabled. 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).[19]

[19]We use 16 occupational codes. The results are similar with a less granular aggregation, or if we replace occupational 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 (7) of Table A.1 in the Appendix.

Table 2: Probit regressions for Type I and Type II errors

	Type I error			*Type II error*		
	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.144***	0.136***	0.128***	-0.037	-0.029	-0.074**
	(0.040)	(0.039)	(0.044)	(0.034)	(0.031)	(0.034)
College degree	0.014	0.016	0.016	0.012	-0.018	-0.027
	(0.040)	(0.040)	(0.040)	(0.035)	(0.034)	(0.034)
Black	-0.019	-0.010	-0.008	-0.083**	-0.095***	-0.114***
	(0.043)	(0.042)	(0.040)	(0.034)	(0.033)	(0.032)
Lab. mark. exp.	-0.006**	-0.005**	-0.004**	0.001	0.000	0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Applied SSI only	-0.153***	-0.157***	-0.138***	0.136***	0.165***	0.138***
	(0.049)	(0.046)	(0.044)	(0.043)	(0.041)	(0.041)
Applied DI + SSI	-0.003	-0.015	-0.018	0.019	0.023	0.030
	(0.047)	(0.047)	(0.046)	(0.044)	(0.042)	(0.040)
Married	0.031	0.031	0.013	0.055	0.090***	0.072**
	(0.042)	(0.041)	(0.039)	(0.036)	(0.033)	(0.033)
Widowed	-0.035	-0.047	-0.080	0.026	0.056	0.054
	(0.072)	(0.068)	(0.066)	(0.062)	(0.060)	(0.058)
Age	-0.014***	-0.015***	-0.015***	0.019***	0.021***	0.022***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Year FE	No	Yes	Yes	No	Yes	Yes
F831 disab. FE	No	Yes	Yes	No	Yes	Yes
HRS Obj. FE	No	No	Yes	No	No	Yes
ADL FE	No	No	Yes	No	No	Yes
BMI+Hosp	No	No	Yes	No	No	Yes
Occupation FE	No	No	Yes	No	No	Yes
[P-value joint sig.]			[0.022]			[0.053]
Observations	772	772	772	833	833	833

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). The "F831 disab. FE" refer to the primary disability codes of SSI/DI applicants (see Table 1); the HRS Obj. FE refer to doctor-diagnosed conditions and the ADL FE refer to indicators for difficulty with activity-of-daily-leaving (see Table B.1 in the Appendix); BMI+Hosp are dummies for extreme BMI levels or hospitalization. The occupational dummies are described in Appendix A.1. ***,**, and * mean significance at 1, 5, and 10 percent level, respectively.

The negative coefficient on the female dummy in columns (4)-(6) of Table 2 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 work disabled and when they are not. However, the estimates are less precise and more unstable across specifications. For these reasons, and because our primary question is on the effectiveness of insurance (as opposed to the moral hazard aspects) of disability insurance, from now on we focus on the evidence for Type I errors.

3.3 Type I Errors at Different Stages of the Evaluation Process

As discussed in Section 2, rejection of an application can occur for different reasons: a "medical" rejection because the health condition is assessed as not being severe and long-lasting; or (sequentially) 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 these data (and the classification explained in Wixon and Strand, 2013) to establish at which stage in the decision process the gender difference in Type I errors arises.

Table 3 breaks down the decision into rejections on health criteria alone ("medical stage") and rejections on the basis of availability of work and residual functional capacity ("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 disability, women are 14.1 percentage points more likely than men to be rejected at the vocational stage.[20]

The clear message is that the difference between men and women in Type I errors arises because of the SSA evaluator's assessments about the ability of applicants 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.

[20] Hu et al. (2001) model the different stages of the SSA disability determination process. Using data from an earlier period (1989-92), they find that rejections for women are significantly higher at the medical stage but insignificantly different from those of men at subsequent stages. Their interpretation is that a higher proportion of women apply with marginal impairments than men. However, in our analysis, our data allows us to compare rejection errors for men and women with the same impairments and self-reported limitations.

Table 3: Type I errors: Rejection at medical or vocational stage

	Medical stage (1)	*Vocational stage* (2)
Female	0.032	0.141***
	(0.033)	(0.047)
All Controls	Y	Y
Sample average	0.162	0.444
Observations	772	647

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. 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) and 0 if rejected at a later stage or awarded. 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) and 0 if awarded. Additional controls are the same as in Table Additional controls are the same as in Table 2.

3.4 Robustness and Extensions

Robustness 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 4. Column (1) reproduces the results of the baseline specification (from Table 2, column (3)). All regressions include the same controls used in the baseline specification.

A first concern is that higher rejection rates for women are due to inherent bias against welfare program recipients, who are typically women (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 slightly 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)-(4) 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 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. The results change very little and are not significantly different from the baseline: among

applicants, women experience higher Type I errors than observationally equivalent men.

Table 4: Probit regressions for Type I errors: Robustness

	Basel.	*DI Sample*	*Timing assumptions*		*Disability definition*	
	(1)	(2)	(3)	(4)	(5)	(6)
			$-2 \leq d \leq 12$	$0 \leq d \leq 12$, weighted	Two ADL+	Predicted
Female	0.128***	0.153***	0.114***	0.118***	0.125***	0.114***
	(0.044)	(0.048)	(0.043)	(0.045)	(0.046)	(0.042)
All Controls	Y	Y	Y	Y	Y	Y
Sample avg.	0.53	0.54	0.54	0.55	0.56	0.57
Obs.	772	586	848	772	622	850

Note: Standard errors in parentheses, clustered at the individual level. The reported coefficients are marginal effects. All regressions include the controls of Table 2, column 3. See notes to that table for a detailed description of the controls. ***, **, and * means significance at 1, 5 and 10 percent level, respectively. Additional controls are the same as in Table 2.

A different set of concerns is about our definition of work disability, which may capture the true work disability status of an individual only imperfectly. In columns (5)-(6) we consider two alternative definitions. In column (5) we assume that an individual is work 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. In column (6) we construct an alternative indicator of work disability in the following way. First, we regress our subjective work disability indicator on clinical and objective indicators and obtain the predictive value, $\hat{p}$. We then define as work disabled those with $\hat{p} \geq E(\hat{p}|Awarded)$. This approach removes one's own interpretation of how objective health conditions affect disability status. The samples and the definitions are different from the baseline, and yet qualitatively the estimates are very similar across specifications, confirming the presence of significant gender differences in Type I errors.

In Table A.1 we consider additional robustness checks. Even if errors are made at the initial evaluation stage, these errors may be 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. The female coefficient rises, but not significantly, from 0.128 (s.e. 0.044) to 0.151 (s.e. 0.045) when we consider the cumulative success rate at the DDS level (initial stage plus potential reconsideration; see column (2)). Unfortunately, we do not have data on outcomes of further appeals (ALJ and beyond). However, even if errors were corrected at later stages, there would still be important

welfare implications from rejections at the DDS level leading to delay. Disability insurance 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.[21]

In column (3) of Table A.1 we restrict the focus to those interviewed up to 9 months following the date of application (instead of 12 as in the baseline). In column (4) of Table A.1 we classify as work 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 work disability used in many papers in the literature and is less strict than our baseline. In column (5) we use an MCA analysis (Hjellbrekke, 2018) exclusively on the clinical/objective disability indicators, extract the first principal inertia, $\hat{a}_i$, and then define as work disabled those with $\hat{a}_i \geq E(\hat{a}_i|Awarded)$. In all cases, the results are qualitatively similar.

Age effects 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 (6) of Table A.1 reports a finer specification with an age spline at these knots. The results are unchanged. The results remain similar (column (7)) if we replace occupational dummies with a physical occupational requirement index using a mapping between HRS occupational codes and O*NET data (as in Michaud and Wiczer, 2018). Applicants with stronger physical requirements at their job are significantly less likely to be turned down if disabled, but women continue to suffer higher Type I error conditioning on that.

Extensions In Table A.2 in the Appendix we conduct a number of specification checks. Column (1) shows the baseline. In column (2), we check if large rejection rates for women are explained by SSA perception of lower economic vulnerability. We interact the female dummy with a dummy for being married and with a dummy for having a spouse with positive earnings. Both interactions are insignificant. Relatedly, 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

[21] 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. Autor, Maestas, Mullen, and Strand (2015) show that additional costs arise from human capital depreciation: longer processing times reduce the employment and earnings of DI applicants for many years after the application, with the effects concentrated among applicants denied benefits at the initial stage.

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 qualitatively similar to the baseline (see column (3)).

A final 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.3) that if anything the interaction suggests *more* persistence among women.

4 Theoretical Framework for Type I Errors

We present a theoretical framework for distinguishing potential explanations for differences in error rates by gender. Suppose that the true, latent work disability (or severe work limitation) status of an individual i is given by:

$$L_i^* = X_i' a_{L^*} + \pi F_i + \varepsilon_i \tag{1}$$

where X_i are observables explaining work limitations (i.e., health variables), F_i is a female dummy, and ε_i represents unobserved heterogeneity in work limitations. The female dummy captures potential shifts in the underlying distribution of work limitations: women may have less severe ($\pi < 0$) or more severe ($\pi > 0$) work impairments than men due for example to genetic differences or behavioral choices. We interpret L_i^* as the measure of work disability targeted by SSA (i.e., the inability to work due to a non-temporary, severe medical condition). This latent work disability status drives both the demand and supply of benefits, which we discuss in turn.

Demand for Benefits

As outlined in Section 3, we use three sequential questions from the HRS to capture as closely as possible the definition of work limitation set by SSA. Nonetheless, self-reports may measure true, latent disability with error. In addition to an individual's perception error about work limitations, our three HRS questions, even if answered without error, could

measure L_i^* imperfectly due to some misalignment with the SSA definition.[22] We define the latent self-reported work disability as:

$$L_i^{**} = L_i^* + \omega_i^i \tag{2}$$

We assume that the noise, ω, does not reflect "justification" bias.[23] Individuals report to be work disabled if their "perceived" work disability status is above an individual threshold: $L_i = \mathbb{1}\left\{L_i^{**} > \bar{L}_i\right\}$, where

$$\bar{L}_i = \bar{L}_{SSA}(X_i) + \gamma F_i + \varphi_i \tag{3}$$

The term $\bar{L}_{SSA}(X_i)$ can be interpreted as the threshold that work limitations would have to cross for an individual to be awarded benefits in a gender-neutral world in which SSA uses only "institutional" variables (such as age, experience or education) to determine eligibility for disability insurance.[24] The individual threshold $\bar{L}_i$ may differ from $\bar{L}_{SSA}(X_i)$ because of random noise (the φ_i term) or because men and women differ in their assessment of the threshold. If, say, women have a lower "pain threshold", $\gamma < 0$, then more women than men will classify themselves as work-disabled despite their underlying L^* being the same, which is the problem of interpersonal comparison of self-reports of work disabilities. This gender difference may also capture heterogeneity in how men and women perceive SSA disability norms (including, in principle, any form of discrimination) or simply reflect potential gender-specific errors in perception.

The individual decision to apply for disability insurance will depend on the expected benefits of applying exceeding the costs:

$$A_i = \mathbb{1}\left\{Q_i E(B_i|R_i = 0) > k_i\right\}$$

where Q_i is the subjective probability of acceptance into the program, B_i the benefit from the program if awarded, and k_i are application costs.

[22] For example, the concept of "persistent disability" in the SSA definition is an impairment that lasts 12 months or more, while the HRS question only distinguishes between impairments that last 3 months or less.

[23] The superscript i indicates this is the work-limitation of individual i, and the subscript indicates that this is evaluated by individual i. This distinction has meaning when we use the vignettes below.

[24] This is an "institutional" form of differentiation because it is embedded in the rules set by SSA for awarding benefits. Besides the fact that applicants must satisfy eligibility criteria (in the case of DI they must have worked in covered employment for a certain number of years and in the case of SSI they must satisfy a means-test), the decision to award benefits assigns more weight to older applicants and applicants with lower levels of education or skills (the so-called "grid", see, e.g., Chen and van der Klaauw, 2008).

Assume that the subjective probability of success is some (monotonically increasing and hence invertible) function $h(.)$ of the distance between perceived work limitations and the individual disability threshold: $Q_i = h(L_i^{**} - \bar{L}_i)$. Hence, the application rule becomes:

$$A_i = \mathbb{1}\left\{ L_i^{**} - \bar{L}_i > h^{-1}\left(\frac{1}{1 + \frac{E(\tilde{B}_i|R_i=0)}{k_i}} \right) \right\}$$

where $\tilde{B}_i = (B_i - k)$ are the net benefits from an award. Assume that $h(.)$ is a logistic function and that the logarithm of the net benefit-cost ratio is a linear function of gender, observables, plus an idiosyncratic term: $\ln \frac{E(\tilde{B}_i|R_i=0)}{k_i} = X_i' a_{\bar{A}} + \tau F_i + v_i$. We can then rewrite the demand for benefits as:

$$A_i = \mathbb{1}\left\{ L_i^{**} > \underbrace{\bar{L}_i + X_i' a_{\bar{A}} + \tau F_i + v_i}_{\bar{A}_i} \right\} \tag{4}$$

The application threshold $\bar{A}_i$ may differ from the "perceived work disability" threshold $\bar{L}_i$ for a number of reasons, even if the two are positively correlated. For example, applicants may "cheat", i.e., apply even when they are not truly work disabled; further, applicants may face different transaction costs of applying, or respond differently to financial incentives to work; finally, 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 this heterogeneity is partly coming from observables (gender and other controls) and partly from unobservables (the term v_i). In particular, the parameter τ 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).

Supply of Benefits

The gender differences in Type I error documented in Section 3.2 may reflect discrimination resulting from the decisions taken by the SSA. A recent literature (Canay et al., 2020) highlights the need to specify the objective function of the decision maker in order to define discrimination in a theoretically coherent way. In keeping with this literature, we outline the formal problem of the SSA examiner and use this to separate institutional differentiation from statistical and taste-based discrimination.

The SSA decides to award or reject applications based on a signal S_i that subsumes both the medical and vocational steps of the assessment process. The signal differs from the true work disability due to random noise ξ_i:

$$S_i = L_i^* + \xi_i \tag{5}$$

We assume that $\xi_i | \left(L_i^*, A_i = 1, F_i\right) \sim N\left(0, \sigma_\xi^2(F_i)\right)$. The variance of the noise can in principle be gender-specific if, for example, women are able to provide more/less accurate documentation about their work limitations than men.

Similarly to Canay et al. (2020) and Arnold et al. (2022), we derive the rejection decision of the SSA examiner from an optimization problem that minimizes total disutility from making screening errors:[25]

$$U = -c_1 \sum_{A_i=1} R_i D_i - c_2 \sum_{A_i=1} \left(1 - R_i\right)\left(1 - D_i\right) \tag{6}$$

where the indicator variable $D_i = \mathbb{1}\{L_i^* \geq \bar{L}_{SSA}(X_i)\}$ identifies individuals with work limitations above a threshold $\bar{L}_{SSA}(X_i)$ that would prevail in a gender-neutral world in which SSA uses only "institutional" variables to determine insurance eligibility.

In the population of applicants, $R_i D_i = 1$ when rejecting a truly work disabled applicant (type I error), which has utility cost c_1, and $\left(1 - R_i\right)\left(1 - D_i\right) = 1$ when awarding benefits to a non-disabled applicant (type II error), which has utility cost c_2. To capture the possibility of utility-based discrimination, we allow the utility cost of Type I and Type II errors to depend on gender, and write them as $c_1(F_i)$ and $c_2(F_i)$ respectively.

Since only a noisy signal of the true work limitations is observed, the problem of SSA examiners consists of minimizing *expected* disutility (6) by choosing whether to turn down ($R_i = 1$) or award ($R_i = 0$) an applicant i. The SSA examiner's expectations are taken with respect to their information set, consisting of observing an applicant's gender F_i and work limitation signal S_i, as well as information on observable characteristics X_i gathered during the application process. In Appendix C, we show that the solution of the examiner's problem is a rejection rule:

$$R_i(X_i, F_i) = \mathbb{1}\left\{p\left(S_i, X_i, F_i\right) < \bar{p}\left(F_i\right)\right\} \tag{7}$$

[25]This is the objective of the SSA evaluators, conditioning on the applications they receive. The objective does not take account of whether the system discourages applicants. We adopt this framework because it captures the focus of our study on screening errors. Further, this mirrors what is often done in other settings where discrimination in "judge" decisions is studied.

where $p\left(S_i, X_i, F_i\right) = E\left(D_i | A_i = 1, S_i, X_i, F_i\right)$ is the examiner's assessed probability that an applicant with characteristics $\{S_i, X_i, F_i\}$ is work limited, and $\bar{p}\left(F_i\right) = \frac{c_2(F_i)}{c_1(F_i)+c_2(F_i)}$ is the marginal assessed probability of work disability triggering an award. This threshold decreases (SSA becomes more lenient) when the utility cost of Type I error increase, and increases (SSA becomes stricter) when the utility cost of Type II error increases.

Equation (7) can be used to distinguish three potential explanations for the gender gap in Type I errors. The first is differences in rejections that arise due to differences in the distribution of "institutional" variables even without explicit statistical or taste-based discrimination. Suppose there is a single such institutional variable X_i, which for simplicity we take to be binary (say, being 55 and above) that increases the probability of success. Institutional gender differences may then be generated if $p(S_i, X_i = 1, F_i) > p(S_i, X_i = 0, F_i)$ and $E(X_i|F_i = 1) < E(X_i|F_i = 0)$. Thus, it is possible that women could suffer on average greater rejections than men with the same signal if their distribution of institutional variables predicting awards has a lower mean than among men. These differences, if they do exist, do not represent "bias" by the decision maker.

The second explanation is statistical discrimination, in which gender is used to improve learning about the unobservables (latent work limitations or unobserved costs of applying). Statistical discrimination against women is present if $p\left(S_i, X_i, F_i = 1\right) < p(S_i, X_i, F_i = 0)$, i.e., if a woman with the same signal and observables as a man is assessed as having a lower probability of being work limited.

Finally, "tastes" for discrimination make the utility losses from Type I error different for men and women, and hence the marginal assessed probability of work disability different. Taste-based discrimination is said to be present if: $\bar{p}(F_i = 1) > \bar{p}(F_i = 0)$, i.e., if women face a higher admission threshold than men. Underlying this notion of taste-based discrimination is that a woman who is otherwise identical to a man has a lower probability of award, whether or not the man and the woman are truly work-limited. In other words, there are lower utility losses from falsely rejecting women (or higher utility losses from falsely awarding them). Note that there is no a priori reason why statistical and taste-based discrimination should go in the same direction and this may mean that positive statistical discrimination may mask negative taste-based discrimination.

The model quantifies the contribution of these three explanations to the estimated gender gap in Type I errors. We rewrite the rule (7), which is cast in terms of $p\left(S_i, X_i, F_i\right)$, in terms of the signal S_i, which is more amenable to the statistical model described below.

To do this, we find the SSA examiner's posterior probability of an applicant being disabled following observation of an applicant's signal, gender and other observables (the examiner's information set). Appendix C shows that this leads to a rejection rule cast in terms of the signal:

$$R_i = \mathbb{1}\left\{S_i < \bar{S}\left(X_i, F_i\right)\right\} \tag{8}$$

where

$$\bar{S}(X_i, F_i) = \frac{\mu_1\left(X_i, F_i\right) + \mu_0\left(X_i, F_i\right)}{2} - \frac{\sigma_S^2(F_i)}{\mu_1\left(X_i, F_i\right) - \mu_0\left(X_i, F_i\right)} \ln\left(\frac{p_D\left(X_i, F_i\right)}{1 - p_D\left(X_i, F_i\right)} \frac{c_1(F_i)}{c_2(F_i)}\right)$$

is the marginal signal triggering an award, μ_1 and μ_0 are the average signal for people with $D_i = 1$ and $D_i = 0$, respectively, and p_D is the probability of being truly work limited conditional on the signal, gender and other observables.[26]

Drivers of Differences in Type I errors

This framework for the demand and supply of disability insurance can be used to perform comparative statics to understand which elements of the model shift the probability of Type I errors, and contribute to explaining gender differences in these errors. In Appendix D we show these comparative statics numerically using our empirical estimates, which (as we discuss in section 5) are obtained imposing joint normality on the distribution of the unobservables. We use the assumption of joint normality, together with the identification of mean and variance shifters by gender, to address the "infra-marginality" problem described in the discrimination literature (Canay et al., 2020), i.e., identification of marginal types from average types. In other contexts, the problem is addressed through quasi-experimental variation in assignment of adjudicator.

Appendix D shows that the probability of Type I error would be higher for women through the following forces:

D1 Women have less severe work limitations ($\pi < 0$);

D2 Women have a lower threshold for reporting a work disability ($\gamma < 0$);

D3 Women have a lower opportunity cost of applying ($\tau < 0$).

[26]Note that we obtain the special case of Arnold et al. (2022) if we impose $\mu_1(.) = 1$ and $\mu_0(.) = 0$.

These parameters have two effects on Type I errors: first, a direct effect from changing the demand for benefits; and second, an indirect effect from shifting the marginal signal triggering an award (the right hand side of (8)).

Two additional (supply) forces explaining gender differences in Type I errors are the noise of the signal $\sigma_S(F_i)$ and the disutility cost of Type I error relative to Type II error, $\frac{c_1(F_i)}{c_2(F_i)}$. From equation (7), $\frac{c_1(F_i)}{c_2(F_i)} = \frac{1-\bar{p}(F_i)}{\bar{p}(F_i)}$, where $\bar{p}\,(F_i)$ is the marginal assessed probability of work disability triggering an award, and hence $\frac{c_1(F_i)}{c_2(F_i)}$ can be interpreted as the marginal odds of an award. Assuming a logistic model, the log of the odds (and of the relative utility cost of Type I error) can be written as a linear function of gender:

$$\ln\left(\frac{c_1(F_i)}{c_2(F_i)}\right) = \alpha_0 + \alpha_1 F_i \tag{9}$$

The parameter α_1 determines the extent of taste-based discrimination in the utility function of the SSA decision maker. The parameter α_0 is normalized to 1 because it cannot be separately identified, so that for men the cost of Type I error is normalised to be the same as the cost of Type II error. In Appendix D we show that the probability of Type I error would be higher for women if:

S1 Women have a noisier disability signal ($\sigma_S(1) > \sigma_S(0)$);

S2 Examiners have a lower disutility cost from making Type I error against women (relative to Type II error) ($\alpha_1 < 0$).

5 Structural Estimates

The framework of Section 4 shows that there are multiple potential drivers of the empirical gender differences in Type I errors. In this section, we first provide a statistical framework corresponding to our theoretical framework. We then outline the identification of the key parameters and corresponding estimates. A common issue in the literature is not knowing the information set of the applicant (or of the defendant in the bail-appeal literature, as in Canay et al., 2020; Arnold et al., 2022). Our empirical context is different from the previous literature because of the availability of self-reports of true work limitation as well as disability vignettes. As discussed below, this additional information is key to our identification.

5.1 Statistical framework

Identification of the structural parameters uses information from five margins: (a) self-reports of work limitations, (b) disability insurance applications, (c) disability vignettes, (d) out-of-pocket health spending, and (e) outcomes of disability insurance applications. The vignettes are important because they help (under some standard restrictions) to pin down inter-personal differences in perceptions of work limitations. Data on out-of-pocket health spending offers a continuous signal of work limitations, which is key for identifying the scale of the various sources of unobserved heterogeneity (subject to at least one normalization). We discuss these five margins in turn.

Self-Reports. Combining Equations (1)-(3) and assuming $\bar{L}_{SSA}(X_i) = X_i' a_{\bar{L}}$, a self-report of work disability L_i is observed if:

$$\begin{aligned} L_i &= \mathbb{1}\left\{L_i^{**} > \bar{L}_i\right\} \\ &= \mathbb{1}\left\{X_i'(a_{L^*} - a_{\bar{L}}) + (\pi - \gamma)F_i + \left(\varepsilon_i + \omega_i^i - \varphi_i\right) > 0\right\} \\ &= \mathbb{1}\left\{X_i'\delta_X^L + \delta_F^L F_i + u_i^L > 0\right\} \end{aligned} \tag{10}$$

Applications. Next, we add information on the decision of whether to apply for disability insurance, combining equations (1)-(4):

$$\begin{aligned} A_i &= \mathbb{1}\left\{L_i^{**} > \bar{A}_i\right\} \\ &= \mathbb{1}\left\{X_i'(a_{L^*} - a_{\bar{L}} - a_{\bar{A}}) + (\pi - \gamma - \tau)\, F_i + \left(\varepsilon_i + \omega_i^i - \varphi_i - v_i\right) > 0\right\} \\ &= \mathbb{1}\left\{X_i'\delta_X^A + \delta_F^A F_i + u_i^A > 0\right\} \end{aligned} \tag{11}$$

If we only have data on self-reports of work disability and disability insurance applications, there is a clear identification problem: it is impossible to separate the effect of (a) genuine worse health from (b) lower pain threshold and (c) lower opportunity costs of applying for disability, given that the scales of (10) and (11) differ. Hence, we cannot assess whether women are more likely to suffer larger Type I error because they have a lower pain threshold ($\gamma < 0$), less severe underlying work limitations ($\pi < 0$), or lower opportunity costs of applying ($\tau < 0$).

Vignettes. To make progress on identification, we use disability vignette data. 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, analogously to our identification problem.

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.[27] 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 you can do?"), and to rank it in terms of severity (possible responses are "None", "Mild", "Moderate", "Severe" and "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.[28] Respondents are asked to rank the vignettes using the same severity scale that was used to rank their own work limitation. To match the analysis from the first part of the paper we convert these responses into a binary indicator, and assume a vignette is assessed as work disabled if the limitation is considered "Severe" or "Extreme". 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 respondents the vignette is labelled as "Mark" and for another respondent the same description refers to a "Tamara".[29]

Since respondents are asked to assess the same broad theoretical construct (work limitations) for themselves and the hypothetical vignette, we assume that the work disability status of the vignette V is governed by the same process as in the general population (equation (1)), i.e.:

$$L_V^* = X_V' a_{L^*} + \pi F_V + \varepsilon_V$$

However, respondent i perceives the work disability status of the vignette V with an i.i.d. error ω_i^V plus, possibly, a "compassion term" capturing the possibility that women respondents classify vignettes as work disabled differently from men respondents (the parameter θ):

[27]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.

[28]As an example, one of the vignettes reads: "[Mark/Tamara] 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 [Mark/Tamara] limited in the kind or amount of work [he/she] could do?".

[29]If the wording of the questions leads respondents to infer the education of the vignette and this changes their opinion about whether the vignette is disabled or not (for example, because respondents associate gender to certain occupations and certain occupations to a higher likelihood of being disabled), then some of the gender differences in disability perceptions may actually reflect occupational differences.

$$L^*_{V,i} = L^*_V + \theta F_i + \omega^V_i$$

Hence, respondent i classifies vignette V as work disabled if:

$$\begin{aligned} L_{V,i} &= \mathbb{1}\left\{L^*_{V,i} > \bar{L}_i\right\} \\ &= \mathbb{1}\left\{X'_V a_{L^*} - X'_i a_{\bar{L}} + (\theta - \gamma) F_i + \pi F_V + \left(\varepsilon_V + \omega^V_i - \varphi_i\right) > 0\right\} \\ &= \mathbb{1}\left\{X'_V \gamma^V_X + X'_i \delta^V_X + \delta^V_F F_i + \gamma^V_F F_V + u^V_i > 0\right\} \end{aligned} \tag{12}$$

where (in the first row) we make the typical assumption made in the vignette literature that respondents use their own threshold $\bar{L}_i$ to assess the presence of work limitations among vignettes ("introspection"). The respondent's gender appears in Equation (12) both because of "compassion" effects as well as "introspection" effects. The vignette's gender appears because the distribution of work limitations among vignettes is assumed to mirror that of the population. The coefficients in Equation 12 can be identified because of the random assignment of vignettes (and vignettes' gender) to respondents.

Out-of-Pocket Spending. Using Equations (10), (11) and (12), we can identify π, γ, τ, and θ but only up to some arbitrary scale. To go beyond this, we introduce a continuous "signal" of work limitations: total out-of-pocket health care spending. In the spirit of common factor structure models, we write (the log of) out-of-pocket spending as:

$$\begin{aligned} C_i &= X'_i a_C + \lambda L^*_i + \psi F_i + \phi_i \\ &= X'_i \left(a_C + a_{L^*}\right) + (\lambda\pi + \psi) F_i + (\lambda\varepsilon_i + \phi_i) \\ &= X'_i \delta^C_X + \delta^C_F F_i + u^C_i \end{aligned} \tag{13}$$

where ψ captures the effect of gender on health care spending over and above the impact through work limitations (e.g., women may be more or less likely to visit doctors independently of work limitations, etc.).

Outcomes of DI/SSI applications. The SSA decision to reject a claim is based on equation (8). The right hand side of (8) is the marginal signal for observing an award and is a general function of observable characteristics and gender, $\bar{S}(X, F)$. We consider a first-order

approximation of this term around $\{\bar{X}_i, F_i = 0\}$ and thus rewrite the rejection rule as:[30]

$$R_i = \mathbb{1}\left\{X_i'\delta_X^R + \delta_F^R F_i + u_i^R > 0\right\} \tag{14}$$

5.2 Estimation Strategy

Equations (10)-(14) are the basis for our identification of the structural parameters. However, to identify the model we need to make distributional assumptions and impose a normalization given that for the most part we deal with binary indicators. Full details are presented in Appendix E. First, we assume that the unobservables in the "self-reporting work disability", "application", "vignette disability reports", "health care spending", and "SSA rejection" decisions (equations (10)-(14)) obey a joint normality assumption:

$$\begin{pmatrix} u_i^L \\ u_i^A \\ u_i^V \\ u_i^C \\ u_i^R \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_L^2 & \rho_{L,A}\sigma_L\sigma_A & \rho_{L,V}\sigma_L\sigma_V & \rho_{L,C}\sigma_L\sigma_C & \rho_{L,R}\sigma_L\sigma_R \\ & \sigma_A^2 & \rho_{A,V}\sigma_A\sigma_V & \rho_{A,C}\sigma_A\sigma_C & \rho_{A,R}\sigma_A\sigma_R \\ & & \sigma_V^2 & \rho_{V,C}\sigma_V\sigma_C & \rho_{V,R}\sigma_V\sigma_R \\ & & & \sigma_C^2 & \rho_{C,R}\sigma_C\sigma_R \\ & & & & \sigma_R^2 \end{pmatrix}\right) \tag{15}$$

where the unobservables relate to the structural error terms:

$$\begin{aligned} u_i^L &= \varepsilon_i + \omega_i^i - \varphi_i \\ u_i^A &= \varepsilon_i + \omega_i^i - \varphi_i - v_i \\ u_i^V &= \varepsilon_V + \omega_i^V - \varphi_i \\ u_i^C &= \lambda\varepsilon_i + \phi_i \\ u_i^R &= -(\varepsilon_i + \xi_i) \end{aligned}$$

We assume that the structural error terms $\varepsilon_i, \omega_i^i, \varphi_i, \omega_i^V, v_i, \phi_i$, and ξ_i are all i.i.d. The variance of ξ_i (noise in the SSA signal) could in principle be gender-specific. However, as we discuss below, there is no evidence of such heteroskedasticity. These assumptions imply that the reduced form unobservables u_i^L, u_i^A, u_i^V, u_i^C, and u_i^R are correlated through sharing

[30] Note that linearity would emerge naturally under the assumption of Arnold et al. (2022) that $\mu_1(.) = 1$ and $\mu_0(.) = 0$ since both $p_D/(1-p_D)$ and c_1/c_2 are odd ratio terms whose logs are linear in the arguments. In practice, a higher-order approximation (where we add quadratic and/or cubic terms in X_i, whenever appropriate, and all the $X_i \times F_i$ interactions) produces almost identical results (in terms of structural parameter estimates and implications). See Appendix G.

common "factors". For example, u_i^L and u_i^V both depend on the term φ_i, while the correlation between u_i^L and u_i^R reflects the unobserved heterogeneity in the true work disability of applicants, ε_i.[31] Besides the joint normality assumption, for identification purposes we also impose two restrictions: (a) $\sigma_\varepsilon^2 = 1$; and (b) $\sigma_{\omega^i}^2 = \sigma_{\omega^V}^2 = \sigma_\omega^2$. Restriction (a) is a normalization, implying that all variances are scaled relative to the variance of unobserved true work disability. Restriction (b) assumes that the errors ω_i^i and ω_i^V (the noise in people's perception of their own work limitations and the noise of people's perception of the vignette's work limitations, respectively) are uncorrelated but have the same variance σ_ω^2. As we shall see, there is not much evidence that these two error terms are correlated.

The goal of the empirical exercise is to estimate the following structural parameters: π, γ, τ, θ, λ, ψ, σ_ω^2, σ_φ^2, σ_v^2, σ_ξ^2, σ_ϕ^2, and α_1. We break estimation in two stages. In the first stage we estimate all parameters except α_1; in the second stage we use the structural parameters from the first stage to pin down α_1. Block bootstrap standard errors correct for this two stage procedure.

The reduced form parameters we use in the first stage are: (i) the coefficient on gender in equations (10)-(13); (ii) the correlation between the various decisions; and (iii) the variance of unobserved health spending (the only decision where the scale can be identified). The mapping from reduced form to structural parameters is obtained using a minimum distance estimator (see Appendix E.3). To obtain the reduced form parameters, we perform maximum likelihood estimation of the parameters describing the joint outcomes of reporting work disability, applying for disability insurance, and being turned down for benefits by SSA conditional on applying (equations (10), (11), and (14)). This is a trivariate probit model with sample selection (since a rejection decision is only observed for applicants). It gives us estimates of parameters κ_1-κ_5 in Table 5, where we show the complete mapping between reduced form and structural parameters. We then perform maximum likelihood estimation (bivariate probit) of the parameters describing the joint outcomes of reporting vignette's and own work disability (equations (12) and (10)).[32] It yields estimates of κ_6-κ_8. Finally, we

[31] It is possible, of course, that the *structural* errors themselves may be correlated. For example, one could argue that people with worse unobserved health would have smaller signal noise, conditioning on observables (i.e., $cov(\varepsilon_i, \xi_i) < 0$). Unfortunately, the latent framework prevents more flexible models to be identified. In Appendix E.6 we show that this particular scenario would produce even stronger evidence for taste-based discrimination. More broadly, our independence assumptions are less strong that they may appear since (besides gender) our regressions control for a very rich set of covariates: education, race, labor market experience, marital status, age, time effects, doctor-diagnosed health conditions, dummies for difficulty with activity of daily living, extreme BMI, hospitalization, occupation, and (in some regressions) dummies for disability codes, health insurance status, and household resources.

[32] We estimate a bivariate probit in order to obtain an estimate of the correlation coefficient between the

estimate the parameters for health spending controlling for censoring at 0 (equations (13)). This gives us the final set of reduced form parameters κ_9-κ_{13}. See Appendix E for more details.[33]

Table 5: The mapping between reduced form and structural parameters

Reduced form parameters	Structural parameters
Applications, Self-reports, Rejection	
$\kappa_1 = \frac{\delta_1^L}{\sigma_L}$	$\frac{(\pi-\gamma)}{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2}}$
$\kappa_2 = \frac{\delta_1^A}{\sigma_A}$	$\frac{(\pi-\gamma-\tau)}{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2+\sigma_v^2}}$
$\kappa_3 = \rho_{L,A}$	$\frac{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2}}{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2+\sigma_v^2}}$
$\kappa_4 = \rho_{A,R}$	$-\frac{1}{\sqrt{1+\sigma_\xi^2}\sqrt{1+\sigma_\omega^2+\sigma_v^2+\sigma_\varphi^2}}$
$\kappa_5 = \rho_{L,R}$	$-\frac{1}{\sqrt{1+\sigma_\xi^2}\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2}}$
Vignettes	
$\kappa_6 = \frac{\delta_1^V}{\sigma_V}$	$-\frac{\gamma-\theta}{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2}}$
$\kappa_7 = \frac{\delta_2^V}{\sigma_V}$	$\frac{\pi}{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2}}$
$\kappa_8 = \rho_{L,V}$	$\frac{\sigma_\varphi^2}{1+\sigma_\omega^2+\sigma_\varphi^2}$
Health Spending	
$\kappa_9 = \delta_1^C$	$\lambda\pi + \psi$
$\kappa_{10} = \sigma_C^2$	$\lambda^2 + \sigma_\phi^2$
$\kappa_{11} = \rho_{C,L}$	$\lambda\frac{1}{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2}}$
$\kappa_{12} = \rho_{C,A}$	$\lambda\frac{1}{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2+\sigma_v^2}}$
$\kappa_{13} = -\rho_{C,R}$	$\lambda\frac{1}{\sqrt{1+\sigma_\xi^2}}$

Given that we have 13 moments (or reduced form parameters) and 11 structural parameters $(\pi, \gamma, \tau, \theta, \lambda, \psi,\ \sigma_\omega^2, \sigma_\varphi^2, \sigma_v^2, \sigma_\xi^2, \sigma_\phi^2)$, we have two overidentifying restrictions that we use to test the model's specification.

unobservables in the two equations ($\rho_{L,V}$, reported at the bottom of Table 6), which we use as moment in the structural estimation exercise. In practice, since this correlation coefficient is close to 0, the estimates obtained with a simple probit for the vignette's work limitations are almost identical.

[33] In principle, one could consider a single maximum likelihood function for the choice margins (10)-(14) using the joint normality assumption (15). However, data on the different margins are not available in all waves. In particular, vignettes are asked only in one wave and the structure of the data is different (an observation is a "respondent/vignette" combination); health spending information is not available in the first two HRS waves and is subject to censoring at 0.

5.3 Identification

Where is identification coming from? Consider the index structure for reporting own disability, applying for benefits, and reporting the disability of a vignette implied by equations (10), (11), and (12) above reproduced here:

$$L_i = \mathbf{1}\left\{ u_i^L > -\left(X_i'(a_{L^*} - a_{\bar{L}}) + \underbrace{(\pi - \gamma)}_{\kappa_1} F_i \right) \right\} \tag{10}$$

$$A_i = \mathbf{1}\left\{ u_i^A > -\left(X_i'(a_{L^*} - a_{\bar{L}} - a_{\bar{A}}) + \underbrace{(\pi - \gamma - \tau)}_{\kappa_2} F_i \right) \right\} \tag{11}$$

$$L_{V,i} = \mathbf{1}\left\{ u_i^V > -\left(X_V' a_{L^*} - X_i' a_{\bar{L}} + \underbrace{(\theta - \gamma)}_{\kappa_6} F_i + \underbrace{\pi}_{\kappa_7} F_V \right) \right\} \tag{12}$$

Ignore for the time being issues of scale being different across indexes. Suppose, for the sake of argument, that in the data woman are more likely than men to report a disability ($\kappa_1 > 0$), apply ($\kappa_2 > 0$), classify a vignette as disabled ($\kappa_6 > 0$), and that respondents are more likely to classify as disabled a female-named than a male-named vignette ($\kappa_7 > 0$). The three equations above have the following interpretation.

From (10), $\kappa_1 > 0$ is consistent with women being genuinely more disabled than men ($\pi > 0$) or having lower pain threshold ($\gamma < 0$). From (11), $\kappa_2 > 0$ may reflect women being truly more disabled than men ($\pi > 0$), having lower pain threshold ($\gamma < 0$), or lower costs of applying ($\tau < 0$). From (12), $\kappa_6 > 0$ is consistent with being more compassionate ($\theta > 0$) or having lower pain threshold themselves ($\gamma < 0$) (since they use introspection when judging how disabled a vignette is, given the verbal description of the condition), while $\kappa_7 > 0$ reflects women being truly more disabled than men in the overall population ($\pi > 0$), since vignettes are interpreted as a draw from the underlying distribution of latent work limitations.

It follows that if we see HRS respondents more likely to classify a female-named vignette as disabled than a male-named vignette with the same condition, we can infer that $\pi > 0$. This pins down π (up to scale), i.e., π is identified by the reduced form parameter κ_7 ($\pi = \boldsymbol{\kappa_7}$). The random assignment of vignette's gender is key, since two identical vignettes will differ only in their male/female assigned name. Once π is known, it is possible to separate excess applications from women due to them being less healthy than men (π) from excess applications due to them having lower pain thresholds ($\gamma = \boldsymbol{\kappa_7 - \kappa_1}$). Once γ is identified,

it is possible to disentangle female compassion from female lower pain threshold in the effect of respondent's gender on the vignette disability reports ($\theta = \boldsymbol{\kappa_6 - \kappa_7 + \kappa_1}$). Finally, once π and γ are identified, the difference between female rate of applications and female rate of disability self reports can be attributed to different opportunity costs of applying ($\tau = \boldsymbol{\kappa_1 - \kappa_2}$).

This basic intuition carries forward accounting for scale differences. In particular, it is straightforward to verify that:

$$\begin{aligned}
\pi &= \boldsymbol{\kappa_7} \left(\sqrt{-\frac{\kappa_{13}}{\kappa_{11}\kappa_5}} \right) \\
\gamma &= (\boldsymbol{\kappa_7 - \kappa_1}) \left(\sqrt{-\frac{\kappa_{13}}{\kappa_{11}\kappa_5}} \right) \\
\theta &= (\boldsymbol{\kappa_6 - \kappa_7 + \kappa_1}) \left(\sqrt{-\frac{\kappa_{13}}{\kappa_{11}\kappa_5}} \right) \\
\tau &= \boldsymbol{\kappa_1} \left(\sqrt{-\frac{\kappa_{13}}{\kappa_{11}\kappa_5}} \right) - \boldsymbol{\kappa_2} \left(\frac{\kappa_5}{\kappa_4} \sqrt{-\frac{\kappa_{13}}{\kappa_{11}\kappa_5}} \right)
\end{aligned}$$

Hence, the same combination of reduced form parameters above ($\kappa_1, \kappa_2, \kappa_6, \kappa_7$) pin down the same structural parameters ($\pi, \gamma, \theta, \tau$), once adjusting for scale effects.[34] In Appendix E we show how the mapping of reduced form parameters identifies the remaining structural parameters, those related to spending ($\lambda, \psi, \sigma_\phi^2$) and the variances ($\sigma_\xi^2, \sigma_v^2, \sigma_\omega^2$, and σ_φ^2).

5.4 Taste-based discrimination parameters

The last part of the identification discussion concerns α_1, the parameter from equation (9) that measures whether the utility cost of Type I error differs by gender (and hence is informative about the presence of taste-based discrimination).

The key equation we use is the SSA rejection rule (8). In Appendix E.4 we show that this rejection rule can be rewritten as:

$$R_i = \mathbf{1}\left\{ \tilde{u}_i^R > -\left(\rho\left(X_i, F_i\right) + \alpha_1 \lambda\left(X_i, F_i\right) \right) \right\} \tag{13}$$

where

[34] Note that the expression for τ accounts for the fact that the application index has a different "scale" that the own or vignette's disability report.

$$\begin{aligned}\rho\left(X_i,F_i\right) &= \frac{m_1\left(X_i,F_i\right)+m_0\left(X_i,F_i\right)}{2\sigma_S}-\frac{\sigma_S}{m_1\left(X_i,F_i\right)-m_0\left(X_i,F_i\right)}\ln\frac{p_D\left(X_i,F_i\right)}{1-p_D\left(X_i,F_i\right)}\\ \lambda\left(X_i,F_i\right) &= -\frac{\sigma_S}{m_1\left(X_i,F_i\right)-m_0\left(X_i,F_i\right)}F_i\end{aligned}$$

and $\tilde{u}_i^R \sim N(0,1)$.

The term $\rho(X_i, F_i)$ reflects statistical discrimination. As we explain in Appendix E.4, estimation of the structural parameters from the first stage (in particular, $\pi, \gamma, \tau, \sigma_\omega^2, \sigma_\varphi^2, \sigma_v^2, \sigma_\xi^2$) allows us to obtain estimates of the terms σ_S, $m_1(X_i, F_i)$, $m_0(X_i, F_i)$, and $p_D(X_i, F_i)$, and thus of $\rho(X_i, F_i)$ and $\lambda(X_i, F_i)$.

Call Π_i the share of applicants with characteristics (X_i, F_i) that are rejected. Since $\tilde{u}_i^R \sim N(0,1)$, we can write:

$$\underbrace{\Pr\left(R_i=1|X_i,F_i\right)}_{\Pi_i}=\Phi\left(\underbrace{\rho\left(X_i,F_i\right)}_{\rho_i}+\alpha_1\underbrace{\lambda\left(X_i,F_i\right)}_{\lambda_i}\right) \tag{14}$$

where $\Phi(.)$ is the standard normal CDF. We can thus identify α_1 using a (scaled) "difference-in-difference" argument:

$$\alpha_1=\frac{\left[E\left(\Phi^{-1}\left(\Pi_i\right)|F_i=1\right)-E\left(\Phi^{-1}\left(\Pi_i\right)|F_i=0\right)\right]-\left[E\left(\rho_i|F_i=1\right)-E\left(\rho_i|F_i=0\right)\right]}{E\left(\lambda_i|F_i=1\right)}$$

In words, when $\alpha_1 = 0$, the last term in (14) disappears and any gender differences in Type I error will reflect only statistical discrimination (if any). This means that the estimate of α_1 is "residual": any excess rejection of female applicants that cannot be attributed to/explained by lower average type ($p_D(X_i, 1) < p_D(X_i, 0)$) or higher average group signal ($m_1(X_i, 1) > m_1(X_i, 0)$) (the terms that produce gender differences in ρ_i) identifies the presence of utility-bias.[35] In terms of implementation, we estimate Π_i using the predicted probabilities from (14).

5.5 Mechanism Parameter Estimates

This section presents the reduced form estimates for the decision to self-report a work disability, apply for DI/SSI, as well as the probability of having a disability insurance

[35]This neglects the possibility that some differences may be explained by "distorted" signals rather than specific utility bias. If signals are distorted, it is impossible to separately identify signal distortions from taste bias (only their combined effect is identified, see Arnold et al., 2022).

claim rejected; the vignette analysis; and out-of-pocket health spending. We then use these estimates as inputs in a minimum distance framework to pin down the mechanism parameters. Appendix E provides details of the maximum likelihood procedure. The key results are presented in Table 6; Appendix F contains the full tables.

5.5.1 Reduced Form Moments

The first three columns of Table 6 report estimates of the parameters affecting the individual's decision to self-report a work disability, to apply for DI/SSI, as well as the decision by the SSA to turn down the application (equations (10), (11), and (14)). In each regression, we use the same controls as in the Type I regressions of Table 2. We use the log of cash-on-hand as an exclusion restriction affecting the decision to apply but not the rejection outcome as this information is not available to DDS examiners and should not enter the screening process.

For the self-reports of work disability, the outcome variable equals 1 if the individual reports to be work disabled; for applications, 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 rejection outcome, 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 disability status.[36] Table 6 (row labeled "Heterosk.(female)") also shows that there is no evidence of heteroskedasticity by gender in any of these three margins (and in the vignette disability reports as well). In what follows we impose homoskedasticity.[37] On a broader level, if the noise of the signal was higher for women, we should expect both Type I and Type II errors to be higher for women than men because women's disability status would be harder to assess. We do not find this to be the case: the gender differences in errors go in opposite directions.

[36] We have fewer observations than in Table 1 because we exclude people with multiple DI/SSI applications in a given year.

[37] This evidence is based on a specific functional form for heteroskedasticity, in which the probability of making choice Y can be written as: $\Pr(Y = 1|X, Z) = \Phi(e^{Zc}X\beta)$. The heteroskedasticity test is a test that $c = 0$ for the female coefficient. In the Appendix, Table A.4, we provide additional support for homoskedasticity that is independent of specific functional form assumptions. In particular, some applicants are rejected because they provide insufficient evidence to support their application. Moreover, some applicants submit evidence that SSA deems in need of supplementation and assigns these applicants to a consultative examination. Both occurrences can be interpreted as a case of a higher noise in the signal initially received by SSA. If we run regressions for having one such occurrence, we find no evidence that women are more likely to submit insufficient or incomplete evidence than men.

Women are slightly more likely to report a work disability and less likely to be applicants, and the effects are significant at least at the 10% level. The fact that women are less likely to apply for disability insurance than men, given self-reported work limitation, is hard to reconcile with the idea that women are more likely than men to "rationalize" employment or DI/SSI participation status with self-reports of a severe work limitation. Finally, the probit for the rejection decision (estimated conditional on applying) confirms the unconditional results from Table 1: women are more likely to have their disability insurance claim rejected.

At the bottom of Table 6 we report the estimated correlation for these three margins. Unobservables in the self-report and application decisions are positively correlated; unobservables in the rejection decisions correlate negatively with those determining self-reports and applications.

Column (4) of Table 6 presents the results of estimating the parameters of equation (12). The dependent variable is whether the respondent classifies a given vignette as disabled. We add controls for the characteristics of the vignette: the vignette's gender and dummies for the vignette health domain. Women respondents are less likely to report that a given vignette is disabled; respondents are less likely to classify a vignette as disabled if that vignette is assigned a female named as opposed to a male name. In a separate specification, we find no evidence that women respondents tend to be more lenient towards female-named vignettes (a coefficient of -0.04 with a s.e. of similar magnitude).

At the bottom of Table 6 we report the estimated correlation between the vignette disability reports and the individual disability self-report. The estimate is small and insignificant (reflecting small heterogeneity in individual disability thresholds).[38]

In the final column (5) of Table 6 we report estimates for the log of out-of-pocket individual medical spending (equation (13)). This includes spending on drugs, hospital in-patient stays, doctor visits, home care, special services and facilities, helpers, nursing homes.[39] In the regression we control for the usual set of variables plus log income, an indicator for low liquid assets (checking account value below $500), and availability of health insurance (private or public). We find that women spend more than men on average. We use

[38] We do not use $\rho_{V,C}$ and $\rho_{V,R}$ because our statistical model sets them to 0, and we do not use $\rho_{A,V}$ because this correlation is very noisy as it is determined only by σ_ψ, which turns out to be insignificant.

[39] A non-negligible share of HRS individuals (about 10%) report exactly zero out-of-pocket health spending. This is partly because most spending may be reimbursed or fully covered by insurance. We account for selection into positive spending using as exclusion restrictions whether there is a spouse that has health insurance coverage, spousal out-of-pocket spending, and dummies for whether the respondent has full coverage of various medical spending. These restrictions are highly jointly significant (p-value $< 0.01\%$). In practice, omitting them does not change the qualitative pattern of the results or their implications.

the residual of this regression to compute the standard deviation of unobserved spending. Further, we compute average unobserved spending conditional on being a work disabled individual, a disability insurance applicant, or a DI/SSI awardee, which gives us estimates of the correlation coefficients between unobserved spending and the self-report, application, and rejection margins. These are reported at the bottom of Table 6, with full details in Appendix E.2.

5.5.2 Mechanism Parameters

Table 7 reports estimates of the mechanism parameters and estimates of the standard deviation of the sources of heterogeneity of the model, using the Minimum Distance mapping between reduced form and structural parameters detailed in Appendix E.3. We calculate standard errors using the Block Bootstrap (based on 200 replications).

The estimates for gender shift of the health distribution, π, and the gender shift of the reporting threshold, γ, are both negative, suggesting in principle some role for health differences and pain threshold perceptions for explaining Type I error differences by gender.[40] By contrast, we find that women appear to have larger opportunity costs of applying than men. "Compassion" effects when assessing vignette work limitations are smaller for women than men. Finally, work disabled individuals have larger out-of-pocket spending and so do women.

The estimates of the standard deviations of unobserved heterogeneity show that individual noise represents about 46% of total variation in self-reports; in contrast, 63% of total unobserved heterogeneity in the signal received by SSA is estimated to be noise. The large estimated noise in the signal received by the SSA aligns with the large estimates of Type I errors in the population (54%). We find no evidence that the individual work disability threshold exhibits much unobserved heterogeneity, while there is large unobserved heterogeneity in the opportunity cost of applying for disability insurance. The final structural estimate is SSA examiners' "taste bias", captured by α_1 which is an estimate of the relative disutility cost of Type I vs Type II errors when assessing women instead of men applicants (see equation (9)). The negative estimate suggests that examiners set a higher admission threshold for women since there is a lower utility cost from incurring Type I errors against

[40] While our estimates suggesting that women have lower 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, with mixed findings. See Racine et al. (2012) and Fillingim et al. (2009) for reviews of this literature.

Table 6: Estimates used for moments: disability self-reports, DI/SSI application, claim rejections, vignettes, out-of-pocket expense

Panel A: Regression Results

	Work lim. self-reports	*Application*	*Claim rejected*	*Vignette work limited*	Log OOP *spending*
	(1)	(2)	(3)	(4)	(5)
Female	**0.041***	**-0.106*****	0.312***	**-0.162*****	**0.187*****
	(0.022)	(0.030)	(0.079)	(0.041)	(0.015)
Female Vignette				**-0.048*****	
				(0.019)	
Log(cash-on-hand)		-0.0156***			
		(0.009)			
σ_C					**1.363*****
					(0.005)
P-value excl. restr.					< 0.001
Heterosk.(female)	0.043	-0.015	-0.275	-0.037	
	(0.027)	(0.040)	(0.178)	(0.033)	
Sample mean	0.05	0.02	0.64	0.37	6.30
Observations	82386	82386	1365	22329	49151

Panel B: Correlations

$\rho_{L,A}$	$\rho_{A,R}$	$\rho_{L,R}$	$\rho_{L,V}$	$\rho_{C,L}$	$\rho_{C,A}$	$\rho_{C,R}$
0.493***	**-0.550*****	**-0.422*****	**0.019**	**0.257*****	**0.155*****	**-0.222****
(0.014)	(0.167)	(0.059)	(0.052)	(0.040)	(0.055)	(0.108)

Note: Reported coefficients in bold are the moments we use in the MD approach below. Marginal effects (when appropriate) are in square brackets, and standard errors in parentheses. Standard errors are clustered at the individual level except standard errors for the correlation coefficients which are obtained with the block bootstrap. ***, **, and * means significance at 1, 5 and 10 percent level, respectively. Controls in all regressions include: education, race, marital status, age, fixed effects for year/wave (except col. (4)), HRS diagnosed health conditions, ADL's, extreme BMI, hospitalization, and occupation. The regression in col. (3) includes, additionally, fixed effects for F831 health code. The regression in col. (4) include fixed effects for vignette's health domain. Finally, exclusion restrictions for col. (5) include whether spouse has insurance, whether spouse has positive spending, amount of spouse spending, and whether individual has full coverage of spending on some health care items. They explain why we observe zero health care expenditure, but not the amount of health care spending conditioning on making some health care purchases. See notes to Table 2 for a detailed description of the controls. See Table F.1 in Appendix F for the full set of results.

Table 7: Structural estimates

Description	Parameter	Estimate
Parameters driving demand		
Health distr. shifter	π	-0.066**
		(0.030)
Disability report threshold	γ	-0.122**
		(0.052)
Application threshold	τ	0.351***
		(0.121)
Compassion effect	θ	-0.345***
		(0.110)
Effect of work disability on spending	λ	0.364***
		(0.113)
Effect of gender on spending	φ	0.211***
		(0.047)
Heterogeneity		
Report noise s.d.	σ_ω	0.929*
		(0.475)
Threshold noise s.d.	σ_φ	0.188
		(0.269)
Application s.d.	σ_v	2.424***
		(0.589)
Spending s.d.	σ_ϕ	1.314***
		(0.025)
Parameters driving supply		
Examiner utility-bias	α_1	-0.641**
		(0.285)
Signal noise s.d.	σ_ξ	1.300***
		(0.469)
OID test statistics		6.59
		[p-value 3.7%]

Note: Diagonally weighted minimum distance standard errors in parenthesis. The standard deviations are relative to the normalisation that the standard deviation of the true latent work limitation, σ_ε is 1.

female applicants.

Since some of the correlation coefficients we use as moments depend on the same structural parameters, the estimation provides two overidentifying restrictions which we use to test the model's specification. Such test, reported in the last row of Table 7, has a p-value of 4%, which is borderline.

6 Implications of Type I Errors

In this final section we consider the implications of Type I errors. In particular: (a) we ask how much of the estimated screening errors can be attributed to observable heterogeneity, and to statistical discrimination vs. utility-based discrimination; (b) we study to what extent are screening error estimates distorted by respondents mis-perceptions of their work limitations or SSA norms; (c) we measure the relative costs of Type I and Type II errors; finally, (d) we consider the labor supply implications of Type I error.

6.1 Distinguishing Statistical Discrimination from Preference based Discrimination

For each individual we compute the predicted Type I error probability as:

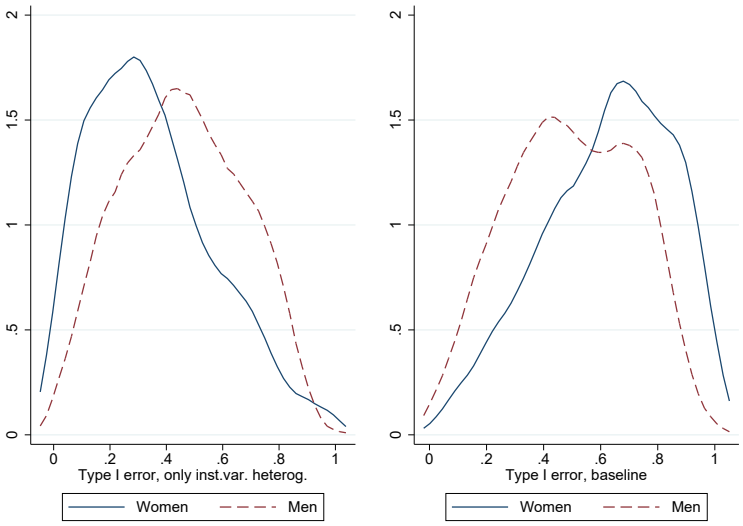
$$\begin{aligned} P_i^{\text{Type I}} &= \Pr(R_i = 1 | A_i = 1, L_i = 1, \boldsymbol{Z}_i) \\ &= \frac{\Phi_3\left(\boldsymbol{Z}_i'\tilde{\delta}^R(\boldsymbol{\beta}), \boldsymbol{Z}_i'\tilde{\delta}^A(\boldsymbol{\beta}), \boldsymbol{Z}_i'\tilde{\delta}^L(\boldsymbol{\beta}); \rho_{L,A}(\boldsymbol{\beta}), \rho_{L,R}(\boldsymbol{\beta}), \rho_{A,R}(\boldsymbol{\beta})\right)}{\Phi_2\left(\boldsymbol{Z}_i'\tilde{\delta}^A(\boldsymbol{\beta}), \boldsymbol{Z}_i'\tilde{\delta}^L(\boldsymbol{\beta}); \rho_{L,A}(\boldsymbol{\beta})\right)} \end{aligned}$$

where $\boldsymbol{Z}_i$ is the vector of individual observations on gender and other observables, and $\tilde{\delta}^j = \frac{\delta^j}{\sigma_J}$ ($j = \{R, A, L\}$) are the ML estimates associated to the decision of rejecting a DI application, the decision to apply, and the decision to self-report a work limitation, respectively (see Table 6). We write these $\tilde{\delta}^j$ and correlation coefficients as functions of $\boldsymbol{\beta}$, the vector of structural parameters of the model. The expression $\Phi_3(.)$ and $\Phi_2(.)$ indicate the trivariate and bivariate normal CDF's, respectively.

We are interested in understanding how changes in the structural parameters affect gender differences in Type I errors:

$$\Delta^{\text{Type I}}(X_i) = E(P_i^{\text{Type I}} | X_i, F_i = 1) - E(P_i^{\text{Type I}} | X_i, F_i = 0)$$

Figure 2: The Distribution of Type I errors



The first row of Table 8, denoted "Estimated Type I Difference", reports the estimate of these differences using the estimated $\tilde{\delta}^j$ $(j = \{R, A, L\})$ and correlation coefficients. The estimate of $\Delta^{\text{Type I}}$ reproduces closely the female-male differentials for Type I errors (13 p.p.) reported in Table 2.

The next three rows of Table 8 decompose this gap into three components: "institutional" differences, statistical discrimination, and taste-based discrimination. We do this by changing the values of key parameters to assess the contribution of these different forces. This in turn changes self-reports, applications and rejection probabilities, and hence the implied Type I error probabilities.

In the second row, rejection rates (and applications and self-reports) differ only because of differences in "institutional" variables that SSA can legitimately use to "differentiate" among applicants (health, age, education, etc.), and there is no discrimination of either sort in the SSA decision. Average Type I error for women are *lower* than for men (0.35 vs 0.46). One explanation is that rejection rates are higher for characteristics that are more common among work-limited men than work-limited women. The increase in Type I errors for men when we allow for observables is driven primarily by the type of health condition that work-limited men are applying with. For example, men are more likely than women to apply with cardiovascular problems and these tend to have high rejection rates.

Average differences may hide considerable heterogeneity. In Figure 2 we plot the

distribution of individual Type I errors separately for men and women under two scenarios: the baseline (right panel) and the counterfactual where only institutional variables are used to determine awards. In other words, all gender effects are shut down and differences in the distribution of Type I errors reflects differences in the distribution of institutional characteristics by gender. The averages in these pictures reproduce the numbers reported in the second row of Table 8. The aspect that emerges clearly from the figure is that accounting for gender effects directly shifts the female distribution from left- to right-skewed, and the mean changes as a consequence.

Table 8: Decomposing Type I Differences

Scenario	All	Women	Men	$\Delta^{\text{Type I}}$
Estimated Type I Difference:	0.57	0.62	0.50	0.12
No discrimination, only institutional var. heterogeneity $\pi = \gamma = \tau = \alpha_1 = 0$	0.39	0.35	0.46	-0.11
Allowing for statistical discrimination: $\alpha_1 = 0$	0.42	0.36	0.50	-0.14
Allowing for taste-based discrimination: $\pi = \gamma = \tau = 0$	0.56	0.60	0.50	0.10

Note: An intermediate case in which we impose no discrimination ($\pi = \gamma = \tau = \alpha_1 = 0$) but allow for all type of observable X heterogeneity (not just institutional variables) generates a Type I error gap of -15 p.p. (35% for women and 50% for men).

The third row of Table 8 allows for all X-heterogeneity and, additionally, for statistical discrimination. However, it maintains the assumption of no taste-based discrimination by setting $\alpha_1 = 0$. Statistical discrimination against women causes a modest increase in the Type I error of 1 percentage points to 0.36. However, the Type I error for women is still below that of men because of the role of observables differing by gender.

The last row allows for utility based discrimination while shutting down statistical discrimination. In this experiment we therefore set $\pi = \gamma = \tau = 0$ which changes the composition of applicants as well as the rejection threshold.[41] Relative to the case which allows only for heterogeneity, taste-based discrimination increases the Type I error for women to 0.60. Comparing this counterfactual with the estimated Type I difference in the first row

[41] One complication is that, since the model is overidentified, the estimated coefficients on the female dummy $\tilde{\delta}^j$ do not necessarily equal their model-implied counterparts ($\tilde{\delta}^j(\hat{\boldsymbol{\beta}})$ (for $j = \{R, L, A\}$). To eliminate the discrepancy and compare counterfactual experiments with the baseline on a neutral basis, we use additive correction factors that are kept constant across experiments.

shows that the gender difference is primarily driven by taste-based discrimination, with only a small role for statistical discrimination.

6.2 Revisiting Type I error rates

An "ideal" estimate of Type I error would condition on true, latent work limitations crossing a gender-neutral threshold that depends only on "institutional variables" and is stripped of any perception error,

$$\Pr(\text{Type I error}) = \Pr(R_i = 1 | A_i = 1, L_i^* > \bar{L}_{SSA}(X_i)) \tag{15}$$

In contrast, we have estimated Type I error rates based on the following:

$$\Pr(\text{Type I error}) = \Pr(R_i = 1 | A_i = 1, L_i^{**} > \bar{L}_i) \tag{16}$$

This is also the standard definition adopted in the literature, see Benitez-Silva et al. (2004) and Low and Pistaferri (2015). The advantage of the definition (16) is that it is immediately measurable, since in our data we observe both the outcome and the conditioning event. From a societal point of view, using an individual-based threshold may be appropriate if people respond truthfully and act on their beliefs.

To understand how much mis-perception errors about own work limitations and about relevant threshold for DI awards contribute to the size of estimated Type I error (and hence to separate supply side "errors" – e.g., screening issues – from demand side "errors" – e.g., norm and work limitation mis-perceptions), we consider two experiments. First, we eliminate gender and unobserved heterogeneity in the perceived threshold. From equation (3) this experiment is equivalent to setting $\gamma = 0, \sigma_\varphi = 0$ and re-defines Type I error as $\Pr(\text{Type I error}) = \Pr(R_i = 1 | A_i = 1, L_i^{**} > \bar{L}_{SSA}(X_i))$, an intermediate step between (15) and (16). The results are reported in Table 9 (row denoted as "$\gamma = 0, \sigma_\varphi = 0$"). Since the estimated γ is small and the extent of heterogeneity in threshold perception (σ_φ) is limited, the results remain very similar to the baseline.

In the second experiment we eliminate noise in self-reported work limitations. If the way people observe their work limitations, or the questions we use from the HRS to measure them, are imperfect, then we might exaggerate the extent of Type I error and possibly the group differences. In particular, our estimates of Type I errors based on Equation (16) treat rejections of applicants that are characterized by $\{L^{**} > \bar{L}_i, L^* < \bar{L}_i\}$ as Type I errors, while in fact rejection is the correct decision by SSA; and treat rejection of applicants with

$\{L^{**} < \bar{L}_i, L^{*} > \bar{L}_i\}$ as correct decisions by the SSA, while in fact they are Type I errors. In our model we estimate the variance of noise in self-reported work disability. This can be used, together with other estimates, to report adjusted estimates of Type I error rates.

This experiment (where we set $\sigma_\omega = 0$) is reported in the penultimate row of Table 9. Type I errors for both genders decline substantially (by about 20 p.p.). However, the gender difference remains basically unchanged (16 p.p. instead of 13 p.p.).

Table 9: Revisiting Type I Errors

Scenario	All	Women	Men	ΔType I
Estimated Type I Difference:	0.57	0.62	0.50	0.13
Gender-independent threshold for self-reports: $\gamma = 0, \sigma_\varphi = 0$	0.57	0.62	0.49	0.13
No Noise in Self-Reports: $\sigma_\omega = 0$	0.38	0.45	0.29	0.16
Gender-independent threshold for self-reports and No Noise in Self-Reports: $\gamma = 0, \sigma_\varphi = 0, \sigma_\omega = 0$	0.37	0.43	0.27	0.16

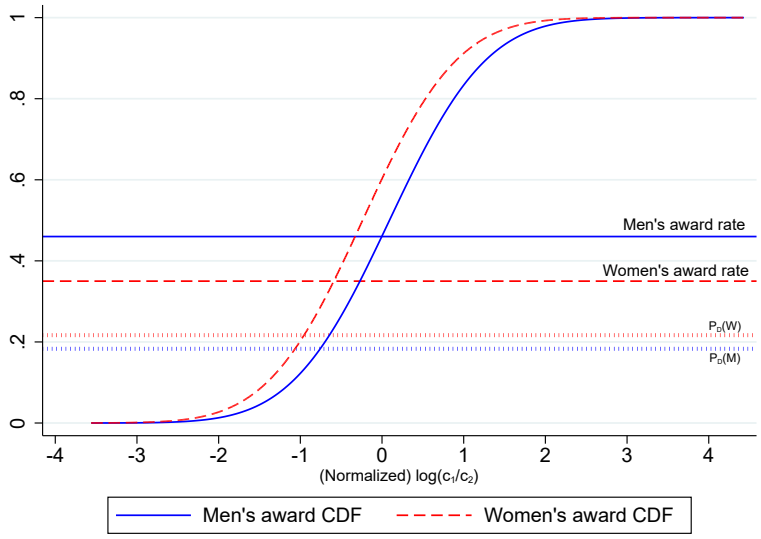
In the final row, we estimate the equivalent of equation (15). This evaluates a work limitation using only institutional variables and without noise in the self-assessment of the work limitation. This confirms that the main driver of the gender difference in Type I error is noise in perceived work limitations, rather than threshold differences.

6.3 The relative costs of Type I vs Type II errors

Our estimate of α_1 suggests that the SSA utility cost of Type I errors (relative to the utility cost of Type II errors) is lower for women than men. However, it does not say whether SSA consider rejecting truly disabled applicants more costly than awarding non-disabled ones.

In Figure 3 we plot the CDF of the award probability (i.e., $\Pr(R_i = 0|A_i = 1)$) against the log of the relative cost of Type I vs Type II error, $\frac{c_1}{c_2}$ (see equation (9)). We do this separately for men and women using the baseline estimates for the remaining parameters. Further, we plot the actual award rate for men (46%) and for women (35%). The point where the CDF crosses the actual award rate pins down the relative cost of screening errors.

Figure 3: Award CDF's and relative cost of screening errors



The crossing for men is normalized at 0.[42]

In the Figure, the actual award rate exceeds the estimates of $p_D(X, F)$, the assessed probability that applicants are disabled. One interpretation is that this reflects the SSA being willing to award benefits to more applicants than would be warranted by its own estimate of applicants' disability because of assigning a greater cost to a Type I versus a Type II error. This is consistent with recent work by Deshpande and Lockwood (2022) arguing that the societal cost of Type II errors are small, since selection into applying for DI/SSI when not severely disabled is mostly by individual hit by severe non-health shocks.

6.4 Labor Market Implications of Type I Errors

In the analysis above, we have inferred the presence of taste-based discrimination from estimation of a structural model. A different approach is to infer discrimination directly from behavior that may be *consequential* to the choice of whether to discriminate. Consider people applying for disability insurance. They can be rejected or awarded ($R = \{1, 0\}$). The unobservable attribute is the extent of their true true work limitations. Assume that if they are truly work limited, they do not work ($P = \{0, 1\}$). Becker (1971) proposes to test for taste-based discrimination using an outcome test. In our setting, this is labor supply

[42]The difference between $\ln\left(\frac{c_1(F=0)}{c_2(F=0)}\right)$ and $\ln\left(\frac{c_1(F=1)}{c_2(F=1)}\right)$ is α_1, and the implied value from the picture is consistent with the estimate reported in Table 7.

decisions after receiving a DI/SSI application rejection. If women who are rejected work less than men who are rejected, the inference is that they were held at higher standards than men.

Table 10 shows the results of these outcome tests. In Test 1, we do not use the self-reports and simply test whether rejected women work more than rejected men. Column (1) shows there is no difference in labour market outcomes by gender in the years after the SSA decision. The problem with interpreting this result is that the group of rejected applicants are a mix of correctly and incorrectly rejected applicants. The mix between these two categories will differ by gender if there are differences in Type I errors or Type II errors. It may also differ if the fraction of applicants who are truly work limited differs by gender.

Test 2 addresses this problem of heterogeneity in the rejected group. If the admission threshold is independent of gender, then the labour supply of incorrectly rejected women and incorrectly rejected men should be the same. However, if the threshold for admission is higher for women than for men, the *average* work limitation of an incorrectly rejected women will be greater than the *average* work limitation of an incorrectly rejected man. If labour supply is a function of the underlying work limitation, we would expect less labour supply among incorrectly rejected women than among incorrectly rejected men.

Column (2) of Table 10 reports results of this test through a three way interaction of being rejected, self-reporting a work limitation and gender. This interaction is significant and negative: incorrectly rejected women work less than incorrectly rejected men. Comparing the results of column (1) and column (2) highlights the value of the additional information on true work limitation status. This information allows us to disentangle rejected applicants to show the outcome for those who have suffered Type I errors.

One concern with the results in column (2) is that individuals who apply for disability and are turned down would not go back to work because of strong preferences for leisure or low permanent productivity. To address this concern, in column (3) we report the results of a regression where the dependent variable is whether the individual was working 5-10 years *prior* to the SSA decision. This is intended as a placebo test to rule out the presence of unobserved heterogeneity driving our results. Controlling for applicants' characteristics, we find that the probability of working appears statistically independent of future reports of work limitation or disability insurance application outcome.

Table 10: Impact of SSA Decision on Subsequent Work

	Test 1	*Test 2*	
	1-3 yrs after	1-3 yrs after	5-10 yrs before
	(1)	(2)	(3)
Female	-0.043*	-0.104***	0.022
	(0.023)	(0.035)	(0.040)
Rejected	0.068**	0.037	-0.027
	(0.027)	(0.040)	(0.035)
Work disabled		-0.085**	0.004
		(0.035)	(0.031)
Rejected × Work disabled		0.036	0.028
		(0.054)	(0.049)
Rejected × Female	0.006	0.129***	0.056
	(0.032)	(0.047)	(0.052)
Work disabled × Female		0.101***	0.040
		(0.040)	(0.047)
Rejected × Work disabled × Female		-0.228***	-0.099
		(0.064)	(0.069)
Observations	1,336	1,336	1,335
Average employment	0.09	0.09	0.78
$R = 1, L = 1$	0.05	0.05	0.77
$R = 1, L = 0$	0.18	0.18	0.75
$R = 0, L = 1$	0.04	0.04	0.84
$R = 0, L = 0$	0.05	0.05	0.76

Note: Standard errors in parentheses, clustered at the individual level. Dependent variable is employment, defined as earning at least as much as the SGA. Respondents are defined as "Work 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. ***, **, and * means significance at 1, 5 and 10 percent level, respectively. Additional controls include: dummies for college degree, Black, marital status, age, labor market experience, whether applied for SSI only, joint SSI/DI application, fixed effects for year, F831 disability code, HRS diagnosed health condition, ADL's, extreme BMI, hospitalization, and occupation. See notes to Table 2 for a detailed description of the controls.

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. Our key conclusion is that gender differences in Type I errors are mostly attributable to taste-based discrimination in the SSA screening process, while only a small fraction can be explained by differences in the distribution of observables or statistical discrimination arising from differences by gender in perceptions of disability or in application costs.

Our results suggest that an important policy change to consider would be to make disability insurance applications gender-blind. 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 three caveats to making the evaluation for disability insurance completely gender blind: first, some disabilities are readily associated with gender; second, the same disability may create different degrees of incapacity by gender, as recent evidence on gender-based medicine suggests; third, evaluators may still use observable characteristics other than gender to infer gender and carry out statistical and taste-based discrimination.

Table 11: Manipulating Information on Gender

Scenario	All	Women	Men
Estimated Type I Difference	0.57	0.62	0.50
No taste-based discrimination	0.42	0.36	0.50
"Ban the (gender) box"	0.56	0.57	0.53

Some of these complex issues are highlighted in Table 11. The first two rows are taken from Table 8, and report the baseline difference in Type I error and the Type I errors that would arise if there was no taste-based discrimination. The 15 percentage point increase in the rate of the Type I error in the overall population is the cost of taste-based discrimination from the point of view of a gender-neutral objective. Since false rejections are more likely to be appealed and this results in more system clogging and delays for all applicants, there is also an indirect cost of the bias. The third row shows a counterfactual experiment in

which we assume that SSA evaluator has no access to information about the applicant's gender. However, the SSA evaluator still has preferences that depend on gender and will use observables to predict gender. This approximates a scenario in which information on gender is withdrawn from the application forms. This scenario is analogous to the "ban-the-box" experiment removing information on criminal history from job application forms (Agan and Starr, 2017). The gender gap in Type I error is much reduced in this experiment, but the overall level of Type I error in the population is little affected.

More broadly, one of the surprising aspects of studying DI/SSI application forms is the type of information that appear on those forms. Some of this information does not appear relevant to assessing the extent of a work disability. For example, the form asks not only about gender, but also about marital status. Further, applicants assessed at the vocational stage are often asked to fill in an additional form, known as Form SSA-3373 (or "Function Report" form). This form contains information on the extent of applicants' residual functional capacity to perform any of their previous jobs or jobs befitting their skills, age, etc.. Several questions included in this form are on activities at home which traditionally are highly gendered.[43] One important issue is whether women are turned down at higher rates than men because they report greater engagement with activities of daily living (ADL) than men in contexts which are traditionally gender-assigned roles. An important question to be addressed in future work is whether the nature of the information collected at these stages may generate the greater rates of rejection observed for women.

[43]For example, "Do you take care of anyone else, such as a wife/husband, children, grandchildren, parents, friend, other?", or "List household chores, both indoors and outdoors, that you are able to do. (For example, cleaning, laundry, household repairs, ironing, mowing, etc.)".

References

Agan, A. and S. Starr (2017, 08). Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment. *The Quarterly Journal of Economics 133*(1), 191–235.

Arnold, D., W. Dobbie, and P. Hull (2022, September). Measuring racial discrimination in bail decisions. *American Economic Review 112*(9), 2992–3038.

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.

Becker, G. (1971). *The Economics of Discrimination* (2 ed.). University of Chicago Press.

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

Bound, J. and R. V. Burkhauser (1999). 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.

Cabral, M. and M. Dillender (2021). Gender Differences in Medical Evaluations: Evidence from Randomly Assigned Doctor. Technical report.

Canay, I. A., M. Mogstad, and J. Mountjoy (2020). On the Use of Outcome Tests for Detecting Bias in Decision Making. NBER Working Papers 27802, National Bureau of Economic Research, Inc.

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.

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.

Deshpande, M. and L. M. Lockwood (2022, July). Beyond health: Non-health risk and the value of disability insurance. *Econometrica 90*(4), 1781–1810.

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.

Hjellbrekke, J. (2018). *Multiple Correspondence Analysis for the Social Sciences*. Routledge.

Hoynes, H. W., N. Maestas, and A. Strand (2022, March). Legal representation in disability claims. Working Paper 29871, National Bureau of Economic Research.

Hu, J., K. Lahiri, D. R. Vaughan, and B. Wixon (2001). A Structural Model of Social Security's Disability Determination Process. *The Review of Economics and Statistics 83*(2), 348–56.

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 (2020). Disability Insurance: Theoretical Trade-Offs and Empirical Evidence. *Fiscal Studies 41*(1), 129–164.

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

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.

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.

Tallis, G. M. (1961). The moment generating function of the truncated multi-normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological) 23*(1), 223–229.

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.

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

Wixon, B. and A. Strand (2013). Identifying SSA's Sequential Disability Determination Steps Using Administrative Data. Research and Statistics Note 2013-01, Office of Retirement and Disability Policy - Office of Research, Evaluation, and Statistics.

A Appendix: Variable Definitions, and Additional Figures and Tables

A.1 Variable Definitions

- A person is classified as having "Some work limitation" if they report to have an impairment or health problem that limits the kind or amount of paid work they can do.

- A person is classified as "Work disabled" if they report to have an impairment or health problem that limits the kind or amount of paid work they can do; if this condition is not temporary (i.e., lasting less than three months); and if the limitation keeps them from working altogether.

- The F831 disability codes categorize applicants according to the primary disability code they apply for: (1) Musculoskeletal, (2) Respiratory, (3) Cardiovascular, (4) Endocrine, (5) Neurological, (6) Mental disorders, (7) Cancer, (8) Immune deficiency, (9) Digestive and Urinary, (10) Others.

- The HRS objective conditions refer to conditions the respondent has been diagnosed with (by a medical provider): high blood pressure, psychological condition, heart condition, arthritis, diabetes, lung condition, stroke, cancer.

- ADL are dummies for difficulties with "activities of daily living": Walking, Dressing, Getting In/Out of Bed, Stooping and crouching.

- The occupational codes are as follows: 1. Managerial specialty operation; 2. Professional specialty operation and technical support; 3. Sales; 4. Clerical, administrative support; 5. Service: protection/Member of Armed Forces; 6. Service: food preparation, private household, cleaning and building services; 7. Health services; 8. Personal services; 9. Farming, forestry, fishing; 10. Mechanics and repairs; 11. Construction, trade and extractors; 12. Precision production; 13. Operators: machine; 14. Operators: transport, etc.; 15. Operators: handlers, etc.; 16. Others/Unknown. Codes 1-9 define the "predominantly female occupations" in Figure 1.

A.2 Additional Figures and Tables

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

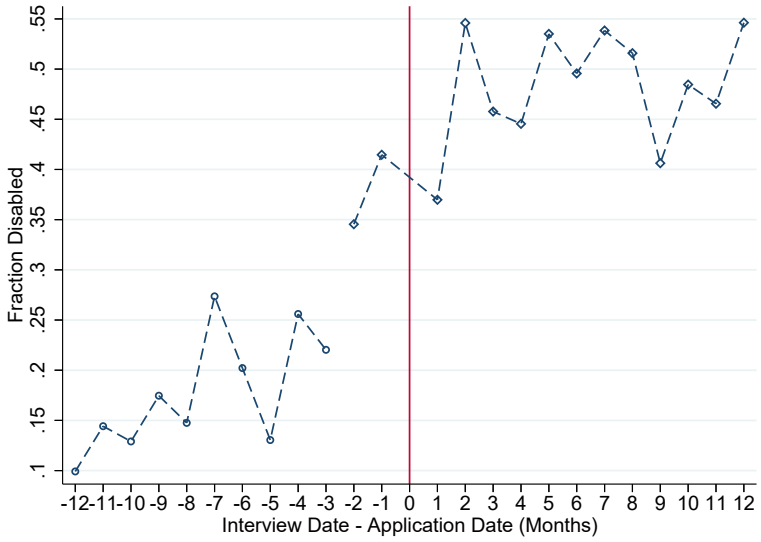


Table A.1: Additional robustness checks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	0.128***	0.151***	0.107**	0.074**	0.108**	0.136***	0.118***
	(0.044)	(0.045)	(0.051)	(0.032)	(0.048)	(0.044)	(0.043)
Age 50-55						-0.050	
						(0.071)	
Age 55-59						-0.255***	
						(0.067)	
Age 60-65						-0.252***	
						(0.071)	
Phys.occ.req. index							-0.010*
							(0.006)
Other demographics	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Health cond. FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
HRS Objective FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ADL FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
BMI+Hosp	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Occupation FE	Yes	Yes	No	Yes	Yes	Yes	No
Sample average	0.53	0.48	0.56	0.58	0.56	0.53	0.53
Observations	772	772	590	1332	706	772	712

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. Column (1) is the baseline. Column (2) is Type I error regression including DDS reconsideration. In column (3) we use those interviewed within 9 months of the application date. In column (4) we classify as work 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 work disability used in many papers in the literature and is less strict than the baseline. In column (5) we use an MCA analysis (Hjellbrekke, 2018) exclusively on the clinical/objective disability indicators, extract the first principal inertia, $\hat{a}_i$, and then define as work disabled those with $\hat{a}_i \geq E(\hat{a}_i|Awarded)$. Column (6) replaces age with an age spline. Column (7) 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).

Table A.2: Additional specification checks

	(1)	(2)	(3)
Female	0.128***	0.122**	0.104**
	(0.044)	(0.060)	(0.044)
Female*empl. husb.		0.064	
		(0.048)	
Female*married		-0.035	
		(0.074)	
Avg. earn. past 5 years			-0.005***
			(0.001)
Primary earn. past 5 years			-0.019
			(0.040)
Other demographics	Yes	Yes	Yes
Health cond. FE	Yes	Yes	Yes
HRS Objective 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	0.53	0.53	0.53
Observations	772	772	772

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. Column (1) is the baseline. Column (2) adds the interaction of the female dummy with married female and a female with an employed husband. Column (3) adds average earnings in the 5 years before application (in real th. $) and a dummy for whether the applicant was the primary household earner over the same period.

Table A.3: Disability transition, ages 20-65

	(1)	(2)	(3)
Disabled at $t-1$	0.385***	0.300***	0.213***
	(0.012)	(0.022)	(0.022)
Female		0.012***	0.010***
		(0.001)	(0.002)
Disabled at $t-1 \times$ female		0.112***	0.099***
		(0.026)	(0.026)
College degree			-0.011***
			(0.002)
Black			0.001
			(0.002)
Married			0.005
			(0.004)
Widowed			0.003
			(0.002)
Age			0.001***
			(0.000)
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	65417	65417	60471

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

Table A.4: Signal Informativeness

	(1)	(2)	(3)
Female	0.002	-0.014	-0.005
	(0.012)	(0.032)	(0.032)
Sample average	0.03	0.49	0.49
Observations	1057	1605	1605

Note: Standard errors in parentheses, clustered at the individual level. The dependent variable in column (1) is whether the applicant was rejected because of insufficient information provided. In columns (2) and (3) the dependent variable is whether the applicant was asked to go through a consultative examination. All regressions control for the same variables used in Table 2, column 3. In column (3) we also add dummies for whether the applicant has health insurance. Marginal effects are reported.

Table A.5: Vignettes: Descriptive Statistics

	All resp.		*Male resp.*		*Female resp.*	
	Mean	SD	Mean	SD	Mean	SD
Vignette disabled						
overall	0.37	0.48	0.39	0.49	0.36	0.48
male hypoth. person	0.37	0.48	0.39	0.49	0.36	0.48
female hypoth. person	0.37	0.48	0.40	0.49	0.35	0.48
Respondent disabled	0.04	0.20	0.03	0.17	0.05	0.21
Respondent's age	57.24	5.06	57.66	4.43	56.98	5.39
Observations	19,143		7,146		11,997	

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.

A.3 Main results using alternative work disability indicators

In this section (Table A.6) we report results using the alternative work disability indicators discussed in the text.

In column (1) we report the baseline. In column (2) the alternative indicator of work disability is constructed in the following way. First, we regress our subjective work disability indicator on clinical and objective indicators and obtain the predictive value, $\hat{p}$. We then define as work disabled those with $\hat{p} \geq E(\hat{p}|Awarded)$. In column (3) the idea is the same, but we run separate regressions for men and women and create the predicted value using only the female coefficients. Finally, in column (4) we use an MCA analysis (Hjellbrekke, 2018) exclusively on the clinical/objective disability indicators, extract the first principal inertia, $\hat{a}_i$, and then define as work disabled those with $\hat{a}_i \geq E(\hat{a}_i|Awarded)$.

We report the relevant marginal effects for five type of regressions: Type I errors (as in column (3) of Table 2), Type II errors (as in column (6) of Table 2), Medical stage and Vocational stage rejections (as in Table 3), and labor supply results (as in column (1) of Table 10).

Table A.6: Using alternative work disability indicators

	(1)	(2)	(3)	(4)
		Type I errors		
Female	0.128***	0.115***	0.116***	0.127***
	(0.044)	(0.042)	(0.043)	(0.049)
		Type II errors		
Female	-0.074**	-0.041	-0.047	-0.072*
	(0.034)	(0.042)	(0.041)	(0.038)
		Medical rejection		
Female	0.032	0.002	0.038	0.051
	(0.033)	(0.033)	(0.032)	(0.039)
		Vocational rejection		
Female	0.141***	0.147***	0.120***	0.097*
	(0.047)	(0.046)	(0.046)	(0.051)
		Labor supply		
Rejected*Work limited	-0.104***	-0.015	-0.042	-0.046
	(0.029)	(0.036)	(0.036)	(0.037)

Note: Specification (1) is the baseline. Specification (2) refers to "Controlling for predicted disability based on the more objective measures"; specification (3) is the same, but the prediction is done using only the coefficients estimated from the male sample; specification (4) uses the MCA approach to create predictions. We report only the (marginal effect) of the female dummy, but all regressions control for the most exhaustive set of controls of the original specification.

B Appendix: Self-reported disability indicators vs. clinical health measures

Self-reported measures of work limitations have pros and cons. On one hand, 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". On the other hand, Bound and Burkhauser (1999) argue that self-reported work disability measures raise three 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 or have stronger preferences for leisure), (b) inter-personal comparability, and (c) perception errors in self-reported disability.

Regarding the first issue, we can verify whether self-reported work disability is associated with more objective or clinical indicators of disability (for which there is less scope for rationalization). We also verify that labor market participation in the years preceding a disability insurance application is not significantly different for people with and without a self-reported work disability. Regarding the second issue, we use responses to disability vignettes to control for person-specific heterogeneity in disability reports (see section 5). Finally, we address the self-perception error issue in various ways: we verify that our results are confirmed when using objective or clinical indicators of disability instead of the self-reported one, and explicitly allow for perceptions errors in our formal model.

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.[44] 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. We also have information on whether a doctor has told the respondents that they have some specific condition, like high blood pressure, cancer, etc. Finally, we have data on individual out-of-pocket health spending.

Table B.1 compares average values of these various health indicators, splitting the sample

[44]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 B.1: Health variables by self-reported work disability status and gender

	Women			*Men*		
	Not work disabled	*Work disabled*	*Diff. (regr.)*	*Not work disabled*	*Work disabled*	*Diff. (regr.)*
	(1)	(2)	(3)	(4)	(5)	(6)
Difficulty walking	0.02	0.17	0.15***	0.01	0.16	0.15***
... dressing	0.03	0.21	0.18***	0.04	0.26	0.22***
... stooping, etc.	0.35	0.80	0.44***	0.27	0.74	0.47***
... getting out of bed	0.03	0.23	0.20***	0.02	0.22	0.19***
... grocery shopping	0.03	0.26	0.23***	0.02	0.20	0.18***
... preparing meals	0.01	0.14	0.13***	0.01	0.09	0.09***
Hospital stay	0.14	0.37	0.23***	0.14	0.44	0.29***
Nights in hospital	0.76	3.55	2.79***	0.87	6.58	5.69***
Obese	0.34	0.47	0.13***	0.31	0.39	0.08***
Underweight	0.01	0.02	0.01**	0.00	0.01	0.01**
Died in sample	0.18	0.39	0.19***	0.24	0.38	0.11***
Doctor diagnosed HBP	0.40	0.64	0.22***	0.44	0.68	0.22***
... psychological condition	0.19	0.46	0.27***	0.10	0.33	0.22***
... heart condition	0.10	0.28	0.17***	0.14	0.34	0.20***
... arthritis	0.46	0.77	0.28***	0.36	0.65	0.28***
... diabetes	0.13	0.27	0.14***	0.15	0.31	0.15***
... lung condition	0.07	0.22	0.15***	0.05	0.17	0.12***
... stroke	0.02	0.09	0.06***	0.03	0.12	0.09***
... cancer	0.08	0.12	0.03***	0.05	0.10	0.05***
Health spending	2404	4965	2514***	1941	5487	3520**

Note: The unit of observation is a person-HRS wave for all variables except death, where it is just person. Respondents are defined as "Work 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.

by self-reports into the "Work disabled" and "Not work disabled" separately for men and for women. Clearly, people who self-report a work disability are much more likely to have a clinical or diagnostic health condition, and more likely to encounter difficulty in ADL's. For example, only about 2% of not disabled women have trouble walking across a room, as opposed to 17% in the disabled group. There are similarly large differences for other ADLs. Mortality for women is 14% vs 31%. Hospital stays are almost three times more likely and five times longer among the work disabled group. Finally, those who self-report a work disability are much more likely to have been diagnosed with a serious health condition and to have larger out-of-pocket health expenditures. 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 B.1 we report the coefficient on the "Work disabled" dummy for each one of the row variables, while controlling for the age of the respondent. The differences are slightly attenuated, but still very sizable and statistically significant throughout. For example, controlling for age, women who self-report to be work disabled have a 15 percentage point higher probability of dying within the period covered by the survey than women who report not to be work disabled. The unconditional difference is 17 percentage points.

Table B.2: Impact of SSA Decision on Subsequent Work

	1-3 yrs after	5-10 yrs before
	(1)	(2)
Rejected	0.097***	-0.001
	(0.023)	(0.029)
Work disabled	-0.032	0.022
	(0.020)	(0.026)
Rejected × Work disabled	-0.104***	-0.029
	(0.029)	(0.037)
Observations	1,256	1,253
Average employment	0.08	0.79
$R = 1, L = 1$	0.03	0.78
$R = 1, L = 0$	0.16	0.76
$R = 0, L = 1$	0.04	0.84
$R = 0, L = 0$	0.05	0.77

Note: Standard errors in parentheses, clustered at the individual level. Dependent variable is employment, defined as earning at least as much as the SGA. Respondents are defined as "Work 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. ***, **, and * means significance at 1, 5 and 10 percent level, respectively. Additional controls include: dummies for gender, college degree, Black, marital status, age, labor market experience, whether applied for SSI only, joint SSI/DI application, fixed effects for year, F831 disability code, HRS diagnosed health condition, ADL's, extreme BMI, hospitalization, and occupation. See notes to Table 2 for a detailed description of the controls.

As mentioned in Section 2.5, to get a gauge of the validity of work limitation self-reports as capturing (albeit with some noise) true screening errors, we test whether applicants who have suffered a Type I error (i.e., rejected applicants who self-report a work disability) work as much as the correctly rejected applicants. This would be consistent with the assessment of the SSA that *all* rejected applicants have residual functional capacity to work. Table B.2, column (1), reports the results of this test, using as dependent variable an indicator for whether the individual had any employment in the three years following the initial consideration stage.[45] We define employment in a given year as the individual having 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.[46]

Our main finding in column (1) is that rejected applicants who self-report a work-limitation work significantly less than those who do not. We interpret this as evidence that Type I errors are true errors: the SSA has overestimated the residual functional capacity of applicants who self-report work-limitations. One concern with the results in column (1) is that individuals who apply for disability and are turned down would not go back to work because of strong preferences for leisure or low permanent productivity. To address this concern, in column (2) we report the results of a regression where the dependent variable is whether the individual was working 5-10 years *prior* to the SSA decision. This is intended as a placebo test to rule out the presence of unobserved heterogeneity driving our results. Controlling for applicants' characteristics, we find that the probability of working appears statistically independent of future reports of work limitation or disability insurance application outcome.

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

[46] This test extends the strategy of Bound (1989) and von Wachter et al. (2011), who compare the labor market outcomes of rejected and non-rejected applicants, by separating further the group of rejected applicants into severely and non-severely work limited applicants.

C Appendix: The Problem of the SSA examiner

We assume that the SSA decides to award or reject applications based on a signal S_i. This signal differs from the true work disability due to a random error:

$$S_i = L_i^* + \xi_i \tag{C.1}$$

where ξ_i is the signal noise. Similarly to Canay et al. (2020) and Arnold et al. (2022), we derive the rejection decision of the SSA examiner from an optimization problem that minimizes disutility from making screening errors. In particular, we express the disutility of making errors of the two types when adjudicating disability insurance applications as:

$$U = -c_1 \sum_{A_i=1} R_i D_i - c_2 \sum_{A_i=1} (1 - R_i)(1 - D_i) \tag{C.2}$$

The indicator variable $D_i = \mathbb{1}\{L_i^* \geq \bar{L}_{SSA}\}$ identifies individuals with work limitations above a gender-neutral threshold $\bar{L}_{SSA}$ that would warrant disability insurance benefits. Hence, in the population of disability insurance applicants, $R_i D_i = 1$ when rejecting a truly work disabled applicant (type I error), which has utility cost c_1, and $(1 - R_i)(1 - D_i) = 1$ when awarding benefits to a non-disabled applicant (type II error), which has utility cost c_2. To capture the possibility of utility-based discrimination, we allow the utility cost of Type I error and of Type II error to depend on gender, and hence express them as $c_1(g)$ and $c_2(g)$, respectively.

The problem of the SSA examiner consists of maximizing expected utility by choosing whether to turn down ($R_i = 1$) or award ($R_i = 0$) an applicant. The SSA examiner chooses $R_i = \{0, 1\}$, but does not observe D_i, so it has to form expectations about it using a disability signal S_i, information about the gender of the applicant F_i (as well as information on observables X_i included in the application form). That is: $E(D_i|S_i, X_i, F_i) = p(S_i, X_i, F_i)$. Formally, the problem becomes:

$$\max_{R_i=\{0,1\}} E(U|S_i, X_i, F_i) = \begin{cases} -c_1(F_i)p(S_i, X_i, F_i) & \text{if } R_i = 1 \\ -c_2(F_i)(1 - p(S_i, X_i, F_i)) & \text{if } R_i = 0 \end{cases}$$

This produces the rejection rule:

$$R_i(S_i, X_i, F_i) = \mathbb{1}\left\{ p(S_i, X_i, F_i) < \underbrace{\left(\frac{c_2(F_i)}{c_2(F_i) + c_1(F_i)} \right)}_{\bar{p}(F_i)} \right\} \tag{C.3}$$

Gender discrimination occurs if $R_i(S_i, X_i, 1) \neq R_i(S_i, X_i, 0)$. This may happen for two reasons: statistical discrimination ($p(S_i, X_i, 1) \neq p(S_i, X_i, 0)$ and/or utility-based discrimination ($\frac{c_1(1)}{c_2(1)} \neq \frac{c_1(0)}{c_2(0)}$). If $p(S_i, X_i, F_i) = p(S_i, X_i)$ we obtain the special case of no statistical discrimination; if the utility costs of screening errors are independent of gender ($\frac{c_1(F_i)}{c_2(F_i)} = \frac{c_1}{c_2}$) we obtain the special case of no taste-based discrimination. There is no a priori reason why statistical and taste-based discrimination should go in the same direction and this may mean that positive statistical discrimination may mask negative taste-based discrimination.

No gender bias is the case in which men and women with identical signals are rejected at identical rates. An increase in the cost of Type I errors makes admissions more lenient; an increase in the cost of Type II errors makes them stricter.

It is convenient to rewrite the rule (C.3), which is cast in terms of $p(S_i, X_i, F_i)$, in terms of the signal S_i, which is more amenable to the statistical model we use. To do this, we find the SSA examiner's posterior probability of an applicant being disabled following observation of an applicant's signal, gender and other observables (the examiner's information set). Using Bayes' Theorem (and omitting from now on the i subscript):

$$\begin{aligned} p(S, X, F) &= \Pr(D = 1|S, X, F) \\ &= \frac{\Pr(S|D = 1, X, F) \Pr(D = 1|X, F)}{\Pr(S|X, F)} \\ &= \frac{q_1(X, F) p_D(X, F)}{q_1(X, F) p_D(X, F) + q_0(X, F)(1 - p_D(X, F))} \end{aligned}$$

The term $p(S, X, F)$ is the (SSA examiner's assessment of the) probability that a person is truly disabled given that it is an applicant of observables X, gender F, "sending" a signal S. It is a function of the following three terms: (1) $p_D(X, F) = \Pr(D = 1|X, F)$, the share of applicants with observables X, gender F who are truly disabled; (2) $q_1(X, F) = \Pr(S|D = 1, X, F)$, the share of truly disabled applicants with observables X, gender F whose observed signal is S; and (3) $q_0(X, F) = \Pr(S|D = 0, X, F)$, the share of non-disabled applicants with observables X, gender F whose observed signal is S.

Note that:

$$\begin{aligned} q_j(X, F) &= \Pr(S|D = j, X, F) \\ &= \frac{1}{\sigma_S} \phi\left(\frac{S - \mu_j(X, F)}{\sigma_S}\right) \end{aligned}$$

because of the normality assumption. Here $\mu_j(X,F) = E(S|D=j,X,F)$ and $\sigma_S = var(S)$ (using the results from the empirical analysis showing that $var(S|.)$ is not a function a gender, or other observables).

It follows that the term $p(S,X,F)$ can be rewritten as:

$$\begin{aligned} p(S,X,F) &= \frac{q_1(X,F)\,p_D(X,F)}{q_1(X,F)\,p_D(X,F) + q_0(X,F)\,(1-p_D(X,F))} \\ &= \frac{\frac{1}{\sigma_S}\phi\left(\frac{S-\mu_1(X,F)}{\sigma_S}\right)p_D(X,F)}{\frac{1}{\sigma_S}\phi\left(\frac{S-\mu_1(X,F)}{\sigma_S}\right)p_D(X,F) + \frac{1}{\sigma_S}\phi\left(\frac{S-\mu_0(X,F)}{\sigma_S}\right)(1-p_D(X,F))} \\ &= \left[1 + e^{-\frac{1}{2\sigma_S^2}\left[2S(\mu_1(X,F)-\mu_0(X,F)) - \left(\mu_1^2(X,F)-\mu_0^2(X,F)\right)\right]}\left(\frac{1-p_D(X,F)}{p_D(X,F)}\right)\right]^{-1} \end{aligned} \tag{C.4}$$

which is strictly increasing in the signal S. Hence, we can substitute this expression in the rejection rule (C.3), and invert $p(S,X,F)$ to get a rule cast in terms of the signal, i.e.

$$R = \mathbf{1}\left\{S < \underbrace{\frac{\mu_1(X,F)+\mu_0(X,F)}{2} - \frac{\sigma_S^2}{\mu_1(X,F)-\mu_0(X,F)}\ln\left(\frac{p_D(X,F)}{1-p_D(X,F)}\frac{1-\bar{p}(F)}{\bar{p}(F)}\right)}_{\bar{S}(X,F)}\right\} \tag{C.5}$$

where $\bar{S}(X,F)$ is the marginal signal for observing an award.

D Appendix: The Effect of Gender on Type I Errors

Section 4 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 SSA signal precision, and differences in how SSA set admission standards. These mechanisms are summarized by the equations below. First, self-reporting a work limitation:

$$\begin{aligned} L_i &= \mathbb{1}\left\{L_i^{**} > \bar{L}_i\right\} \\ &= \mathbb{1}\left\{X_i'(a_{L^*} - a_{\bar{L}}) + (\pi - \gamma)F_i + \left(\varepsilon_i + \omega_i^i - \varphi_i\right) > 0\right\} \end{aligned} \tag{D.1}$$

Next, the decision of whether to apply for disability insurance:

$$\begin{aligned} A_i &= \mathbb{1}\left\{L_i^{**} > \bar{A}_i\right\} \\ &= \mathbb{1}\left\{X_i'(a_{L^*} - a_{\bar{L}} - a_{\bar{A}}) + (\pi - \gamma - \tau)\, F_i + \left(\varepsilon_i + \omega_i^i - \varphi_i - v_i\right) > 0\right\} \end{aligned} \tag{D.2}$$

These are equations (10) and (11) in the main text.

The SSA decision is:

$$\text{Reject if: } S_i \;<\; \bar{S}(X_i, F_i) \tag{D.3}$$

where, as explained in Section 4, the threshold $\bar{S}(X_i, F_i)$ reflects statistical- and taste-based discrimination (see equation (C.5)).

We use these equations to characterize Type I errors. We then simulate how Type I errors vary with each of the key parameters: $\{\pi, \gamma, \tau, \alpha_1\}$, respectively representing gender differences in underlying health, pain thresholds, application thresholds, and gender differences in SSA relative utility costs of Type I vs Type II errors.

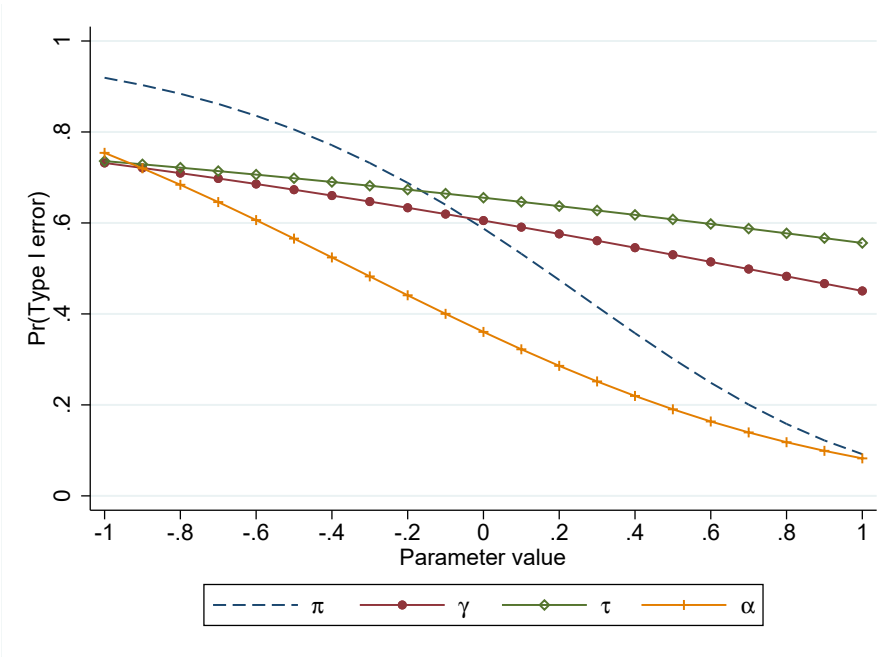
We define the Type I error as:

$$\begin{aligned} \Pr\left(R_i = 1 | L_i = 1, A_i = 1, X_i\right) &= \Pr\left(S_i < \bar{S}(X_i, F_i) | L_i^{**} > \bar{L}_i, L^{**} > \bar{A}_i, X_i\right) \\ &= \Pi\left(\pi, \gamma, \tau, \alpha_1, \boldsymbol{\delta}\right) \end{aligned}$$

where $\boldsymbol{\delta}$ is the vector of other structural parameters besides the ones of interest.

If we make a joint normality assumption about the unobservables $(\varepsilon_i, \omega_i, \varphi_i, v_i, \xi_i)$, we can simulate how changes in one of the parameters of interest (say, π) affects the Type I

Figure D.1: Simulated Type I Errors



error probability while keeping all the other parameters $(\gamma, \tau, \alpha_1, \boldsymbol{\delta})$ constant. We use the empirical estimates in Table 7 to perform this exercise. The results are shown in Figure D.1. The Type I error probability for women declines as they become more truly work disabled $(\frac{\partial \Pi(.)}{\partial \pi} < 0)$; as the threshold for classifying themselves as work-disabled increases $(\frac{\partial \Pi(.)}{\partial \gamma} < 0)$; as their implied application threshold increases $(\frac{\partial \Pi(.)}{\partial \tau} < 0)$; and as the examiner's utility cost of making a Type I error against them (as opposed to a male applicant) increases $(\frac{\partial \Pi(.)}{\partial \alpha_1} < 0)$.

E Appendix: Identification

E.1 Reduced Form Equations and Error Structure

Our model includes an equation for the decision to report a disability:

$$\begin{aligned} L_i &= \mathbb{1}\left\{L_i^{**} > \bar{L}_i\right\} \\ &= \mathbb{1}\left\{X_i'(a_{L^*} - a_{\bar{L}}) + (\pi - \gamma)F_i + \left(\varepsilon_i + \omega_i^i - \varphi_i\right) > 0\right\} \\ &= \mathbb{1}\left\{X_i'\delta_X^L + \delta_F^L F_i + u_i^L > 0\right\} \\ &= \mathbb{1}\left\{\boldsymbol{Z}_i'\tilde{\delta}^L + \epsilon_i^L > 0\right\} \end{aligned} \tag{E.1}$$

where $\boldsymbol{Z}_i = [\, X_i \quad F_i \,]$ and $\epsilon_i^L \sim N(0,1)$. Here and below, $\tilde{\delta}^j = \frac{\delta^j}{\sigma_j}$ and δ^j is the vector of relevant (pre-scaled) parameters.

Next, we add information on the decision of whether to apply for disability insurance:

$$\begin{aligned} A_i &= \mathbb{1}\left\{L_i^{**} > \bar{A}_i\right\} \\ &= \mathbb{1}\left\{X_i'(a_{L^*} - a_{\bar{L}} - a_{\bar{A}}) + (\pi - \gamma - \tau)\, F_i + \left(\varepsilon_i + \omega_i^i - \varphi_i - v_i\right) > 0\right\} \\ &= \mathbb{1}\left\{X_i'\delta_X^A + \delta_F^A F_i + u_i^A > 0\right\} \\ &= \mathbb{1}\left\{\boldsymbol{Z}_i'\tilde{\delta}^A + \epsilon_i^A > 0\right\} \end{aligned} \tag{E.2}$$

with $\epsilon_i^A \sim N(0,1)$.

We use information on the work disabilities of vignettes, as assessed by HRS respondents:

$$\begin{aligned} L_{V,i} &= \mathbb{1}\left\{L_{V,i}^* > \bar{L}_i\right\} \\ &= \mathbb{1}\left\{X_V' a_{L^*} - X_i' a_{\bar{L}} + (\theta - \gamma)\, F_i + \pi F_V + \left(\varepsilon_V + \omega_i^V - \varphi_i\right) > 0\right\} \\ &= \mathbb{1}\left\{X_V'\gamma_X^V + X_i'\delta_X^V + \delta_F^V F_i + \gamma_F^V F_V + u_i^V > 0\right\} \\ &= \mathbb{1}\left\{\boldsymbol{Z}_{Vi}'\tilde{\delta}^V + \epsilon_i^V > 0\right\} \end{aligned} \tag{E.3}$$

where and $\epsilon_i^V \sim N(0,1)$ and $\boldsymbol{Z}_{Vi} = [\, X_V \quad X_i \quad F_i \quad F_V \,]$.

We add information on out-of-pocket health care spending:

$$\begin{aligned} C_i &= X_i' a_C + \lambda L_i^* + \psi F_i + \phi_i \\ &= X_i'\left(a_C + a_{L^*}\right) + (\lambda\pi + \psi)\, F_i + (\lambda\varepsilon_i + \phi_i) \\ &= X_i'\delta_X^C + \delta_F^C F_i + u_i^C \end{aligned} \tag{E.4}$$

Finally, we use outcomes of disability insurance applications (i.e., whether a claim is rejected):

$$\begin{aligned} R_i &= \mathbb{1}\{S_i < \bar{S}(X_i, F_i)\} \\ &\approx \mathbb{1}\left\{X_i'\delta_X^R + \delta_F^R F_i + u_i^R > 0\right\} \\ &= \mathbb{1}\left\{\boldsymbol{Z}_i'\tilde{\delta}^R + \epsilon_i^R > 0\right\} \end{aligned} \tag{E.5}$$

with $\epsilon_i^R \sim N(0,1)$.[47]

As reported in the main text, we assume that the unobservables in the "self-reporting work disability", "application", "SSA rejection", "vignette disability reports", and "health care spending" decisions obey a joint normality assumption:

$$\begin{pmatrix} u_i^L \\ u_i^V \\ u_i^A \\ u_i^C \\ u_i^R \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_L^2 & \rho_{L,V}\sigma_L\sigma_V & \rho_{L,A}\sigma_L\sigma_A & \rho_{L,C}\sigma_L\sigma_C & \rho_{L,R}\sigma_L\sigma_R \\ & \sigma_V^2 & \rho_{V,A}\sigma_A\sigma_R & \rho_{V,C}\sigma_V\sigma_C & \rho_{V,R}\sigma_V\sigma_R \\ & & \sigma_A^2 & \rho_{A,C}\sigma_A\sigma_C & \rho_{A,R}\sigma_A\sigma_R \\ & & & \sigma_C^2 & \rho_{C,R}\sigma_C\sigma_R \\ & & & & \sigma_R^2 \end{pmatrix}\right)$$

where, to recall, the error terms have the following structure:

$$\begin{aligned} u_i^L &= \varepsilon_i + \omega_i^i - \varphi_i \\ u_i^V &= \varepsilon_v + \omega_i^V - \varphi_i \\ u_i^A &= \varepsilon_i + \omega_i^i - \varphi_i - v_i \\ u_i^C &= \lambda\varepsilon_i + \phi_i \\ u_i^R &= -(\varepsilon_i + \xi_i) \end{aligned}$$

For identification purposes, we assume that the structural error terms $\varepsilon_i, \omega_i^i, \varphi_i, \omega_i^V, v_i, \phi_i$, and ξ_i are all i.i.d. (the variance of ξ could be gender-specific in principle; in practice, there is no evidence of heteroskedasticity by gender). These assumptions imply that the reduced form unobservables u_i^L, u_i^V, u_i^A, u_i^C, and u_i^R are correlated through sharing common "factors". For example, u_i^L and u_i^A both depend on the composite term $\varepsilon_i + \omega_i^i - \varphi_i$, while

[47] As we explain in the main text, we consider a first-order approximation of the term $\bar{S}(X_i, F_i)$ around $\{\bar{X}_i, F_i = 0\}$. A second order approximation produces almost identical results.

the correlation between u_i^L and u_i^R reflects the unobserved heterogeneity in the true work disability of applicants, ε_i.

To identify the parameters of interest we use moments reflecting the five "choices" above: self-reporting a work disability, reporting a vignette work disability, applying for disability insurance, the amount of out-of-pocket health care spending, and the SSA disability insurance application rejection outcome (conditioning on an application being observed). Our distributional assumptions (together with the functional form assumption of the various indexes) allow us to estimate "reduced form" parameters and hence the mapping to the structural parameters of interest (detailed below) under two restrictions: (a) $\sigma_\varepsilon^2 = 1$ (normalization), implying that all variances are scaled relative to the variance of unobserved true work disability; (b) $\sigma_{\omega^i}^2 = \sigma_{\omega^V}^2 = \sigma_\omega^2$ (the noise in people's perception of their own work disability has the same variance as the noise of people's perception of the vignette's work disability). This restriction has a simple interpretation: the errors ω_i^i and ω_i^V represent noise in the respondent's assessment of work disability (respectively, noise about their own work limitations and about the vignette's work disability). We assume that these errors are uncorrelated but that they have the same variance σ_ω^2.

The goal is to estimate the following structural parameters: π, γ, τ, θ, λ, ψ, σ_ω^2, σ_φ^2, σ_v^2, σ_ξ^2, σ_ϕ^2, and α_1. We break estimation in two stages. In the first stage we estimate all parameters except α_1; in the second stage we use the structural parameters estimated in the first stage to pin down α_1. Block bootstrap standard errors correct for this two stage procedure.

E.2 Moments

A first set of moments come from considering the joint outcomes of: (a) reporting work disability (b) applying for disability insurance, and (c) being turned down for benefits by SSA conditional on applying (equations (E.1), (E.2), and (E.5)). To reduce cluttering we omit the observable characteristics X_i below, but they are fully accounted for in estimation. The reduced form parameters that are estimated and the mapping with the structural parameters are as follows:

Reduced form par.	Structural par.
$\kappa_1 = \frac{\delta_1^L}{\sigma_L}$	$\frac{(\pi-\gamma)}{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2}}$
$\kappa_2 = \frac{\delta_1^A}{\sigma_A}$	$\frac{(\pi-\gamma-\tau)}{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2+\sigma_v^2}}$
$\kappa_3 = \rho_{L,A}$	$\frac{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2}}{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2+\sigma_v^2}}$
$\kappa_4 = \rho_{A,R}$	$-\frac{1}{\sqrt{1+\sigma_\xi^2}\sqrt{1+\sigma_\omega^2+\sigma_v^2+\sigma_\varphi^2}}$
$\kappa_5 = \rho_{L,R}$	$-\frac{1}{\sqrt{1+\sigma_\xi^2}\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2}}$

We obtain estimates of parameters κ_1-κ_5 by maximum likelihood estimation of a trivariate probit model with sample selection (since a rejection decision is only observed for applicants):

$$
\begin{aligned}
L \;=\; & \prod_{i=1}^{n} \Pr\left(A_i=0, L_i=1; \kappa_1,\kappa_2,\kappa_3\right)^{(1-A_i)L_i} \Pr\left(A_i=L_i=0; \kappa_1,\kappa_2,\kappa_3\right)^{(1-A_i)(1-L_i)} \\
& \Pr\left(A_i=L_i=R_i=1; \kappa_1,\kappa_2,\kappa_3,\kappa_4,\kappa_5\right)^{A_iL_iR_i} \Pr\left(A_i=L_i=1, R_i=0; \kappa_1,\kappa_2,\kappa_3,\kappa_4,\kappa_5\right)^{A_iL_i(1-R_i)} \\
& \Pr\left(A_i=R_i=1, L_i=0; \kappa_1,\kappa_2,\kappa_3,\kappa_4,\kappa_5\right)^{A_i(1-L_i)R_i} \\
& \Pr\left(A_i=1, L_i=R_i=0; \kappa_1,\kappa_2,\kappa_3,\kappa_4,\kappa_5\right)^{A_i(1-L_i)(1-R_i)} \qquad \text{(E.6)}
\end{aligned}
$$

Next, we use moments referring to the joint decision of reporting a vignette's disability and own disability:

Reduced form par.	Structural par.
$\kappa_6 = \frac{\delta_1^V}{\sigma_V}$	$-\frac{\gamma-\theta}{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2}}$
$\kappa_7 = \frac{\delta_2^V}{\sigma_V}$	$\frac{\pi}{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2}}$
$\kappa_8 = \rho_{L,V}$	$\frac{\sigma_\varphi^2}{1+\sigma_\omega^2+\sigma_\varphi^2}$

and obtain estimates of κ_6-κ_8 using bivariate probit.[48]

Finally, we have moments involving out-of-pocket health care spending (conditional on observing positive spending, which is measured by the indicator $I_i^{C>0}$). From (E.4), we have the following mapping:

[48] In principle, one could embed the disability vignette responses with the three other decision (reporting own disability, apply, being rejected). However, disability vignette data are only available for a single HRS wave and the structure of the data is different (an observation is a "respondent/vignette" combination). We thus obtain the parameters $\kappa_6-\kappa_8$ come from a bivariate probit for reporting own disability and the vignette's disability. We estimate a bivariate probit in order to obtain an estimate of the correlation coefficient between the unobservables in the two equations ($\rho_{L,V}$, reported at the bottom of Table 6), which we use as moment in the structural estimation exercise. In practice, since this correlation coefficient is close to 0, the estimates obtained with a simple probit are almost identical.

Reduced form par.	Structural par.
$\kappa_9 = \delta_1^C$	$\lambda\pi + \psi$
$\kappa_{10} = \sigma_C^2$	$\lambda^2 + \sigma_\phi^2$

as well as average unobserved out-of-pocket health care spending conditioning on reporting non-zero spending and (respectively) reporting a disability, applying for benefits, and being awarded benefits:

Reduced form par.	Structural par.
$\kappa_{11} = \rho_{C,L}$	$\lambda\frac{1}{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2}}$
$\kappa_{12} = \rho_{C,A}$	$\lambda\frac{1}{\sqrt{1+\sigma_\omega^2+\sigma_\varphi^2+\sigma_v^2}}$
$\kappa_{13} = -\rho_{C,R}$	$\lambda\frac{1}{\sqrt{1+\sigma_\xi^2}}$

where κ_{11}, κ_{12} and κ_{13} are the coefficients of a regression of the moments $E\left(u_i^C | I_i^{C>0} = 1, L_i = 1\right), E\left(u_i^C | I_i^{C>0} = 1, A_i = 1\right)$, and $E\left(u_i^C | I_i^{C>0} = 1, R_i = 0, A_i = 1\right)$, respectively, onto the generalized inverse Mills ratio appropriate for the selected sample. In particular, one can show that, for generic random variables x_j and generic truncation points t_j:[49]

$$E\left(x_1 | x_2 > t_2, x_3 > t_3\right) = \rho_{12}\phi\left(t_2\right)\frac{\Phi\left(-\frac{t_3-\rho_{23}t_2}{\sqrt{1-\rho_{23}^2}}\right)}{\Phi_2\left(-t_2, -t_3; \rho_{23}\right)} + \rho_{13}\phi\left(t_3\right)\frac{\Phi\left(-\frac{t_2-\rho_{23}t_3}{\sqrt{1-\rho_{23}^2}}\right)}{\Phi_2\left(-t_2, -t_3; \rho_{23}\right)}$$

and

$$\begin{aligned} E\left(x_1 | x_2 > t_2, x_3 > t_3, x_4 > t_4\right) &= \rho_{12}\phi\left(t_2\right)\frac{\Phi_2\left(-\frac{t_3-\rho_{32}t_2}{\sqrt{1-\rho_{32}^2}}, -\frac{t_4-\rho_{42}t_2}{\sqrt{1-\rho_{42}^2}}; \rho_{34.2}\right)}{\Phi_3\left(-t_2, -t_3, -t_4; \rho\right)} \\ &+\rho_{13}\phi\left(t_3\right)\frac{\Phi_2\left(-\frac{t_2-\rho_{23}t_3}{\sqrt{1-\rho_{23}^2}}, -\frac{t_4-\rho_{43}t_3}{\sqrt{1-\rho_{43}^2}}; \rho_{24.3}\right)}{\Phi_3\left(-t_2, -t_3, -t_4; \rho\right)} \\ &+\rho_{14}\phi\left(t_4\right)\frac{\Phi_2\left(-\frac{t_2-\rho_{24}t_4}{\sqrt{1-\rho_{24}^2}}, -\frac{t_3-\rho_{34}t_4}{\sqrt{1-\rho_{34}^2}}; \rho_{23.4}\right)}{\Phi_3\left(-t_2, -t_3, -t_4; \rho\right)} \end{aligned}$$

in which $x_j \sim N\left(0, 1\right)$, $\rho_{ij} = corr(x_i, x_j)$, and

$$\rho_{xy.z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{1-\rho_{xz}^2}\sqrt{1-\rho_{yz}^2}}$$

[49] We use formulae from Tallis (1961).

We use estimates of the trivariate probit for self-reporting work disability, application, and disability insurance awards, to construct the generalized Mills' ratio terms in the expression above. We then regress the health spending residual onto these terms to recover estimates of the relevant parameters κ_{11}-κ_{13} for the appropriately selected samples (work-limited; disability insurance applicants; awarded applicants – all with positive health spending).

In the main text we show how the combination of these moments identify some of the structural parameters:

$$\begin{aligned}
\pi &= \boldsymbol{\kappa_7}\left(\sqrt{-\frac{\kappa_{13}}{\kappa_{11}\kappa_5}}\right) \\
\gamma &= (\boldsymbol{\kappa_7 - \kappa_1})\left(\sqrt{-\frac{\kappa_{13}}{\kappa_{11}\kappa_5}}\right) \\
\theta &= (\boldsymbol{\kappa_6 - \kappa_7 + \kappa_1})\left(\sqrt{-\frac{\kappa_{13}}{\kappa_{11}\kappa_5}}\right) \\
\tau &= \boldsymbol{\kappa_1}\left(\sqrt{-\frac{\kappa_{13}}{\kappa_{11}\kappa_5}}\right) - \boldsymbol{\kappa_2}\left(\frac{\kappa_5}{\kappa_4}\sqrt{-\frac{\kappa_{13}}{\kappa_{11}\kappa_5}}\right)
\end{aligned}$$

Now we show how the mapping of reduced form parameters identifies the remaining structural parameters, those related to spending $(\lambda, \psi, \sigma_\phi^2)$ and the variances $(\sigma_\xi^2, \sigma_v^2, \sigma_\omega^2$, and $\sigma_\varphi^2)$. In particular, one can show that the spending parameters are identified by:

$$\begin{aligned}
\lambda &= \sqrt{-\frac{\kappa_{11}\kappa_{13}}{\kappa_5}} \\
\psi &= \kappa_9 + \frac{\kappa_{13}\kappa_7}{\kappa_5} \\
\sigma_\phi^2 &= \kappa_{10} + \frac{\kappa_{11}\kappa_{13}}{\kappa_5}
\end{aligned}$$

and finally, the variances of the remaining unobservables (normalized with respect to to the variance of true work disability unosbervables, $\sigma_\varepsilon^2 = 1$):

$$
\begin{aligned}
\sigma_{\xi}^{2} &= -\left(1+\frac{\kappa_{11}}{\kappa_{5}\kappa_{13}}\right) \\
\sigma_{v}^{2} &= \frac{\kappa_{13}}{\kappa_{11}\kappa_{5}}\left(1-\left(\frac{\kappa_{5}}{\kappa_{4}}\right)^{2}\right) \\
\sigma_{\omega}^{2} &= -\left(1+(1-\kappa_{8})\frac{\kappa_{13}}{\kappa_{11}\kappa_{5}}\right) \\
\sigma_{\varphi}^{2} &= -\kappa_{8}\frac{\kappa_{13}}{\kappa_{11}\kappa_{5}}
\end{aligned}
$$

Given that we have 13 moments (or reduced form parameters) and 11 structural parameters $(\pi, \gamma, \tau, \theta, \lambda, \psi,\ \sigma_{\omega}^{2}, \sigma_{\varphi}^{2}, \sigma_{v}^{2}, \sigma_{\xi}^{2}, \sigma_{\phi}^{2})$, we have two overidentifying restrictions that can be used to test the model's specification.

E.3 Appendix: Minimum Distance Estimation

The estimates of the structural parameters of the model are obtained using a simple Minimum Distance procedure. The previous section has illustrated the mapping between reduced form parameters κ_j $(j = 1...13)$ and the structural parameters. Call $\boldsymbol{\beta} = \{\pi, \gamma, \tau, \theta, \lambda, \psi,$ $\sigma_{\omega}^{2}, \sigma_{\varphi}^{2}, \sigma_{v}^{2}, \sigma_{\xi}^{2}, \sigma_{\phi}^{2}\}$ the vector of structural parameters, $\hat{\boldsymbol{\kappa}}$ the vector of estimated reduced form parameters, and $\boldsymbol{\kappa}(\boldsymbol{\beta})$ the theoretical mapping between reduced form and structural parameters. The structural estimates $\boldsymbol{\beta}$ are obtained by solving:

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

where $\boldsymbol{\Omega}$ is a weighting matrix. Optimal minimum distance (OMD) sets $\boldsymbol{\Omega} = \boldsymbol{V}^{-1}$, where $\boldsymbol{V}$ is the variance matrix of $\hat{\boldsymbol{\kappa}}(\boldsymbol{\beta})$, which we obtain by the block bootstrap (with 200 replications). To avoids the pitfalls of optimal minimum distance (OMD) remarked by Altonji and Segal (1996), which are primarily related to the terms outside the main diagonal of the optimal weighting matrix, we set $\boldsymbol{\Omega}$ to be the diagonal of $\boldsymbol{V}^{-1}$, which gives the diagonally-weighted minimum distance (DWMD) estimator. Unlike the equally weighted minimum distance (EWMD) estimator, DWMD allows for heteroskedasticity.[50] The standard errors of the DWMD estimates $\hat{\boldsymbol{\beta}}$ are obtained from the variance matrix:

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{G}'\boldsymbol{\Omega}\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{\Omega}\boldsymbol{V}\boldsymbol{\Omega}\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{\Omega}\boldsymbol{G})^{-1}$$

[50] Using EWMD produces very similar estimates and almost identical implications regarding the role of institutional differences vs. statistical and taste-based discrimination.

where $\boldsymbol{G}$ is is the Jacobian matrix evaluated at the estimated parameters. The overidentification test is based on the statistic:

$$L = (\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}(\boldsymbol{\beta}))'(\boldsymbol{W}\text{var}(\hat{\boldsymbol{\beta}})\boldsymbol{W}')^{-}(\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}(\boldsymbol{\beta}))$$

where $\boldsymbol{W} = \mathbf{I} - \boldsymbol{A}$, $\boldsymbol{A} = \boldsymbol{G}(\boldsymbol{G}'\boldsymbol{\Omega}\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{\Omega}$ and $^{-}$ indicates a generalized Moore-Penrose inverse. Under the null, $L \sim \chi^2_{oid}$.

E.4 Recovering α_1

In this section we show how we recover an estimate of α_1, which is key to the distinction between statistical- and utility-based discrimination. Recall the SSA rejection rule (C.5), reproduced here (again omitting for simplicity the i subscript):

$$R = \mathbf{1}\left\{ S < \underbrace{\frac{\mu_1(X,F) + \mu_0(X,F)}{2} - \frac{\sigma_S^2}{\mu_1(X,F) - \mu_0(X,F)} \ln\left(\frac{p_D(X,F)}{1 - p_D(X,F)} \frac{c_1(F)}{c_2(F)} \right)}_{\bar{S}(X,F)} \right\} \tag{E.1}$$

Since we find no evidence that the variance of the signal varies by gender, we impose $\sigma_S^2(F) = \sigma_S^2$. We now derive expressions for $p_D(X,F)$ and $\mu_j(X,F)$ ($j = \{0,1\}$) that depend on the structural parameters of the model.

E.4.1 The term $p_D(X,F)$

Call D an indicator for being "truly" disabled, i.e., $D = \mathbb{1}\left\{L^* \geq \bar{L}_{SSA}(X)\right\}$, where $\bar{L}_{SSA}(X)$ is the SSA gender-neutral threshold that depends only on "institutional" factors such as age, experience, education, etc. (estimated from disability self-reports). Since $L^* = X'a_{L^*} + \pi F + \varepsilon$, the event $\mathbb{1}\{D = 1\}$ corresponds to $\mathbb{1}\{\varepsilon > -(X'(a_{L^*} - a_{\bar{L}}) + \pi F)\}$. From now on, define $(X'(a_{L^*} - a_{\bar{L}}) + \pi F) = c^D(X,F)$. Hence, using (E.2),

$$\begin{aligned} p_D(X,F) &= \Pr(D = 1 | A = 1, X, F) \\ &= \Pr\left(\varepsilon > -c^D(X,F) \,|u^A > -\left(X'\delta_X^A + \delta_F^A F\right)\right) \\ &= \frac{\Pr\left(\varepsilon > -c^D(X,F), u^A > -\left(X'\delta_X^A + \delta_F^A F\right)\right)}{\Pr\left(u^A > -\left(X'\delta_X^A + \delta_F^A F\right)\right)} \\ &= \frac{\Phi_2\left(c^D(X,F), c^A(X,F); \rho_{\varepsilon,u^A}\right)}{\Phi\left(c^A(X,F)\right)} \end{aligned} \tag{E.2}$$

where $c^A(X,F) = \frac{X'\delta_X^A + \delta_F^A F}{\sigma_A}$, $\Phi_2(.)$ is the bivariate standard normal CDF, and $\rho_{\varepsilon,u^A} = \left(1+\sigma_\omega^2+\sigma_\varphi^2+\sigma_v^2\right)^{-\frac{1}{2}}$. We can thus use the estimates of our structural parameters to obtain an estimate of $p_D(X,F)$.

E.4.2 The terms $\mu_1(X,F)$ and $\mu_0(X,F)$

The term $\mu_1(X,F)$ is:

$$\begin{aligned}
\mu_1(X,F) &= E(S|D=1,A=1,X,F) \\
&= E(L^*+\xi|D=1,A_i=1,X,F) \\
&= E(X'a_{L^*}+\pi F+\varepsilon+\xi|D=1,A=1,X,F) \\
&= E(X'a_{L^*}+\pi F|D=1,A=1,X,F)+E(\varepsilon|D=1,A=1,X,F) \\
&= (X'a_{L^*}+\pi F)+\phi\left(c^D(X,F)\right)\frac{\Phi\left(\frac{c^A(X,F)-\rho_{\varepsilon,u^A}c^D(X,F)}{\sqrt{1-\rho^2_{\varepsilon,u^A}}}\right)}{\Phi_2\left(c^D(X,F),c^A(X,F);\rho_{\varepsilon,u^A}\right)} \\
&\quad +\rho_{\varepsilon,u^A}\phi\left(c^A(X,F)\right)\frac{\Phi\left(\frac{c^D(X,F)-\rho_{\varepsilon,u^A}c^A(X,F)}{\sqrt{1-\rho^2_{\varepsilon,u^A}}}\right)}{\Phi_2\left(c^D(X,F),c^A(X,F);\rho_{\varepsilon,u^A}\right)} \\
&= (X'a_{L^*}+\pi F)+m_1(X,F) \qquad \text{(E.3)}
\end{aligned}$$

This can be interpreted as the average signal observed for truly disabled applicants with observable characteristics X and gender F. Symmetrically, the term $\mu_0(X,F)$ is:

$$\begin{aligned}
\mu_0(X,F) &= E(S|D=0,A=1,X,F) \\
&= (X'a_{L^*}+\pi F)-\phi\left(c^D(X,F)\right)\frac{\Phi\left(\frac{c^A(X,F)-\rho_{\varepsilon,u^A}c^D(X,F)}{\sqrt{1-\rho^2_{\varepsilon,u^A}}}\right)}{\Phi_2\left(-c^D(X,F),c^A(X,F);-\rho_{\varepsilon,u^A}\right)} \\
&\quad +\rho_{\varepsilon,u^A}\phi\left(c^A(X,F)\right)\frac{1-\Phi\left(\frac{c^D(X,F)-\rho_{\varepsilon,u^A}c^A(X,F)}{\sqrt{1-\rho^2_{\varepsilon,u^A}}}\right)}{\Phi_2\left(-c^D(X,F),c^A(X,F);-\rho_{\varepsilon,u^A}\right)} \\
&= (X'a_{L^*}+\pi F)+m_0(X,F) \qquad \text{(E.4)}
\end{aligned}$$

which is the average signal observed for non-disabled applicants of observable characteristics X and gender F.

E.4.3 The modified SSA rejection rule

Replacing (E.2), (E.3), and (E.4) into (E.1), and using the functional form assumption for $\frac{c_1(F)}{c_2(F)}$ (equation (9)), gives the modified SSA rejection rule:

$$R = \mathbf{1}\left\{\tilde{u}^R > -\left(\rho\left(X,F\right) + \alpha_1 \lambda\left(X,F\right)\right)\right\} \tag{E.5}$$

where

$$\begin{aligned}
\rho\left(X,F\right) &= \frac{m_1\left(X,F\right) + m_0\left(X,F\right)}{2\sigma_S} - \frac{\sigma_S}{m_1\left(X,F\right) - m_0\left(X,F\right)} \ln \frac{p_D\left(X,F\right)}{1 - p_D\left(X,F\right)} \\
\lambda\left(X,F\right) &= -\frac{\sigma_S}{m_1\left(X,F\right) - m_0\left(X,F\right)} F
\end{aligned}$$

and $\tilde{u}^R \sim N(0,1)$.

Call Π the share of applicants with characteristics (X,F) that are rejected. Since $\tilde{u}^R \sim N(0,1)$, we can write:

$$\underbrace{\Pr\left(R=1|X,F\right)}_{\Pi} = \Phi\left(\underbrace{\rho\left(X,F\right)}_{\rho} + \alpha_1 \underbrace{\lambda\left(X,F\right)}_{\lambda}\right) \tag{E.6}$$

where $\Phi(.)$ is the standard normal CDF. We can thus identify α_1 using a (scaled) "difference-in-difference" argument:

$$\alpha_1 = \frac{\left[E\left(\Phi^{-1}\left(\Pi\right)|F=1\right) - E\left(\Phi^{-1}\left(\Pi\right)|F=0\right)\right] - \left[E\left(\rho|F=1\right) - E\left(\rho|F=0\right)\right]}{E\left(\lambda|F=1\right)}$$

In words, when $\alpha_1 = 0$, the second term on the righ-hand side of (E.6) disappears and any gender differences in Type I error will reflect statistical discrimination (if any). This means that the estimate of α_1 is "residual": any excess rejection of female applicants that cannot be attributed to/explained by lower average type ($p_D(X,1) < p_D(X,0)$) or higher average group signal ($m_1(X,1) > m_1(X,0)$) (the terms that produce gender differences in ρ) identifies the presence of taste-bias.[51] In terms of implementation, we do not have a non-parametric estimate of Π_i (because the sample of applicants is small and the X_i set is large), and hence use the predicted probabilities from the probit estimation of equation (14).

[51] This neglects the possibility that some differences may be explained by "distorted" signals rather than specific utility bias. If signals are distorted, it is impossible to separately identify signal distortions from utility bias (only their combined effect is identified, see Arnold et al., 2022).

E.5 The special case of no statistical discrimination

As discussed in Section 4, the adjudicator's assessed probability of true work disability is given by:

$$p\left(S,X,F\right) = \left[1+e^{-\frac{1}{2\sigma_S^2}\left[2S(\mu_1(X,F)-\mu_0(X,F))-\left(\mu_1(X,F)^2-\mu_0(X,F)^2\right)\right]}\left(\frac{1-p_D\left(X,F\right)}{p_D\left(X,F\right)}\right)\right]^{-1} \quad \text{(E.7)}$$

where we have used the fact that the empirical analysis shows no evidence for gender-specific heteroskedasticity in the signal noise.

We now show that $p(S,X,F) = p(S,X)$, i.e., there is no statistical discrimination by gender, if we impose the parametric restrictions $\pi = \gamma = \tau = 0$.

Consider first the term $p_D(X,F)$. From section E.4.1, this is:

$$p_D(X,F) = \frac{\Phi_2\left(c^D\left(X,F\right), c^A\left(X,F\right); \rho_{\varepsilon,u^A}\right)}{\Phi\left(c^A\left(X,F\right)\right)}$$

where:

$$c^D(X,F) = X'(a_{L^*} - a_{\bar{L}}) + \pi F$$

and, from equation (E.2),

$$c^A\left(X,F\right) = \frac{X'(a_{L^*} - a_{\bar{L}} - a_{\bar{A}}) + (\pi - \gamma - \tau)F}{\sigma_A}.$$

Imposing the restrictions $\pi = \gamma = \tau = 0$ makes $p_D(X,F)$ independent of gender and only a function of observable characteristics X.

Next, consider the terms $\mu_1(X,F)$ and $\mu_0(X,F)$. From section E.4.2, it is again clear that the restrictions $\pi = \gamma = \tau = 0$ make both terms independent of gender and only a function of observable characteristics X.

Since $p_D(X,F)$, $\mu_1(X,F)$ and $\mu_0(X,F)$ are all independent of gender when $\pi = \gamma = \tau = 0$, it follows that, in equation (E.7), $p(S,X,F) = p(S,X)$ as well.

E.6 Bias from ignoring correlation among structural errors

Starting from (E.6) and defining $y_i = \Phi^{-1}\left(\Pi_i\right)$ to simplify notation, we have:

$$y_i = \rho_i + \alpha_1 \lambda_i$$

from which we obtain:

$$\hat{\alpha}_1 = \frac{\left[E\left(y_i|F_i=1\right) - E\left(y_i|F_i=0\right)\right] - \left[E\left(\rho_i|F_i=1\right) - E\left(\rho_i|F_i=0\right)\right]}{E\left(\lambda_i|F_i=1\right)}$$

To identify α_1 we use the empirical estimates of ρ_i and λ_i. Assume now that $cov\left(\varepsilon,\xi\right) = \sigma_{\varepsilon\xi} \neq 0$. In particular, one might expect $\sigma_{\varepsilon\xi} < 0$ (if applicants are more work limited there is a smaller noise in the signal). The true variance of the unobservable component of the signal is:

$$\sigma_S^2 = var\left(\varepsilon + \xi\right) = 1 + \sigma_\xi^2 + 2\sigma_{\varepsilon\xi}$$

To construct estimates of ρ_i and λ_i we are instead using

$$\begin{aligned}
\hat{\sigma}_S^2 &= 1 + \hat{\sigma}_\xi^2 \\
&= -\frac{\kappa_{11}}{\kappa_5\kappa_{13}} \\
&= \frac{\left(1 + \sigma_\xi^2 + 2\sigma_{\varepsilon\xi}\right)}{\left(1+\sigma_{\varepsilon\xi}\right)^2} = \theta^2\sigma_S^2
\end{aligned}$$

where $\theta^2 = \frac{1}{\left(1+\sigma_{\varepsilon\xi}\right)^2}$. Hence, we would also be estimating ρ_i and λ_i with bias because both terms depend on σ_S^2. In particular, we can rewrite the terms that we use to identify α_1 as:

$$\begin{aligned}
\hat{\rho}_i &= \frac{m_1\left(X,F\right) + m_0\left(X,F\right)}{2\theta\sigma_S} - \theta\sigma_S\frac{1}{m_1\left(X,F\right) - m_0\left(X,F\right)}\ln\frac{p_D\left(X,F\right)}{1-p_D\left(X,F\right)} \\
&= \frac{a_i}{\theta\sigma_S} - \theta\sigma_S b_i c_i \\
\hat{\lambda}_i &= -\theta\sigma_S\frac{1}{m_1\left(X,F\right) - m_0\left(X,F\right)}F \\
&= -\theta\sigma_S b_i F_i \\
y_i &= \frac{a_i}{\sigma_S} - \sigma_S b_i c_i - \sigma_S\alpha_1 b_i F_i
\end{aligned}$$

where $a_i = m_1\left(X_i,F_i\right) + m_0\left(X_i,F_i\right)$, $b_i = \left(m_1\left(X_i,F_i\right) - m_0\left(X_i,F_i\right)\right)^{-1}$, $c_i = \ln\frac{p_D(X_i,F_i)}{1-p_D(X_i,F_i)}$.

We have unbiased estimates of the terms a_i, b_i, c_i and of $(\theta\sigma_S)$ but not of θ or σ_S separately. However, we can show that:

$$p\lim\hat{\alpha}_1 = \frac{\alpha_1}{\theta} + (1-\theta)\frac{E\left(a_i|F=1\right) - E\left(a_i|F=0\right)}{\left(\theta\sigma_S\right)^2 E\left(b_i|F=1\right)} + \frac{(1-\theta)}{\theta}\frac{E\left(b_i c_i|F=1\right) - E\left(b_i c_i|F=0\right)}{E\left(b_i|F=1\right)}$$

For values of $\sigma_{\varepsilon\xi} \leq 0$ (and using our estimates of the terms a_i, b_i, c_i and of $(\theta\sigma_S)$), the estimate of α_1 is downward biased, i.e., the evidence in favor of taste-based discrimination would be even stronger.

F Reduced Form Estimates Used as Moments: Full Results

In this Appendix we report the full results of Table 6.

Table F.1: Full ML estimates of Table 6

	Self-reports	Application	Claim rejected	Vignette	Log spending	Spending> 0
Female	0.041	-0.106	0.312	-0.162	0.187	0.349
	(0.022)	(0.030)	(0.079)	(0.041)	(0.015)	(0.021)
College degree	-0.148	-0.079	0.136	-0.064	0.045	0.167
	(0.020)	(0.029)	(0.077)	(0.038)	(0.015)	(0.020)
Black	0.101	0.183	0.166	0.267	-0.130	-0.170
	(0.022)	(0.029)	(0.094)	(0.055)	(0.021)	(0.023)
Lab. mark. experience	-0.019	0.005	0.003	-0.001	-0.002	0.004
	(0.001)	(0.001)	(0.004)	(0.002)	(0.001)	(0.001)
Doctor diagnosed HBP	0.149	0.140	-0.689	0.080	0.373	0.538
	(0.046)	(0.067)	(0.216)	(0.061)	(0.023)	(0.031)
Doctor diagnosed psych. cond.	0.565	0.573	-0.737	0.235	0.544	0.401
	(0.049)	(0.068)	(0.217)	(0.076)	(0.034)	(0.048)
Doctor diagnosed heart cond.	0.508	0.468	-0.607	0.083	0.626	0.448
	(0.053)	(0.076)	(0.230)	(0.107)	(0.036)	(0.053)
Doctor diagnosed arthritis	0.435	0.412	-0.735	0.032	0.392	0.437
	(0.032)	(0.048)	(0.160)	(0.048)	(0.018)	(0.024)
Doctor diagnosed diabetes	0.589	0.529	-0.546	0.086	0.729	0.715
	(0.036)	(0.053)	(0.176)	(0.064)	(0.024)	(0.032)
Doctor diagnosed lung disease	0.830	0.681	-0.822	-0.017	0.658	0.574
	(0.039)	(0.056)	(0.184)	(0.089)	(0.032)	(0.045)
Doctor diagnosed stroke	0.762	0.679	-1.243	-0.054	0.733	0.495
	(0.048)	(0.066)	(0.189)	(0.153)	(0.046)	(0.058)
Doctor diagnosed cancer	0.535	0.533	-0.754	0.024	0.607	0.550
	(0.040)	(0.058)	(0.186)	(0.068)	(0.027)	(0.040)
Married	-0.078	-0.122	0.087	-0.034	0.003	0.180
	(0.021)	(0.028)	(0.079)	(0.043)	(0.025)	(0.027)
Widowed	0.019	-0.103	-0.176	-0.026	0.000	0.000
	(0.035)	(0.051)	(0.136)	(0.083)	(.)	(.)
Age	0.021	-0.029	-0.050	0.010	0.015	0.005
	(0.002)	(0.002)	(0.012)	(0.003)	(0.001)	(0.002)
ADL, diff. walking	0.444	0.170	-0.634	-0.121	0.260	-0.086
	(0.035)	(0.046)	(0.104)	(0.155)	(0.050)	(0.059)
ADL, diff. dressing	0.314	0.335	-0.317	0.045	0.172	-0.087
	(0.030)	(0.039)	(0.091)	(0.114)	(0.036)	(0.044)
ADL, diff. stooping etc	0.590	0.365	-0.184	-0.046	0.196	0.046
	(0.019)	(0.028)	(0.096)	(0.038)	(0.015)	(0.021)
ADL, diff. getting out of bed	0.398	0.316	0.056	0.144	0.191	0.054
	(0.030)	(0.039)	(0.104)	(0.112)	(0.040)	(0.050)
Some nights in hosp.	0.422	0.475	-0.488	-0.050	0.752	0.594
	(0.019)	(0.026)	(0.076)	(0.048)	(0.018)	(0.031)
BMI	-0.001	-0.004	-0.011	0.004	0.000	0.004
	(0.001)	(0.002)	(0.005)	(0.003)	(0.001)	(0.002)
Occ.: Unknown	-0.297	-0.141	0.357	-0.117	-0.018	0.213
	(0.035)	(0.053)	(0.137)	(0.083)	(0.028)	(0.036)
Occ.: Manag/Prof	-0.145	-0.055	0.508	0.000	-0.039	0.161
	(0.032)	(0.049)	(0.137)	(0.085)	(0.028)	(0.036)
Occ.: Sales/clerical	-0.094	0.008	0.088	0.024	-0.119	0.007
	(0.039)	(0.055)	(0.141)	(0.103)	(0.036)	(0.041)
Occ.: Clean/Protect/Food serv	-0.122	0.028	0.049	-0.073	-0.096	0.053
	(0.037)	(0.053)	(0.136)	(0.098)	(0.035)	(0.041)
Occ.: Personal/Health serv	0.021	0.027	0.160	-0.141	-0.083	0.033
	(0.042)	(0.059)	(0.151)	(0.099)	(0.035)	(0.041)
Occ.: Farm/Constr/Mech	0.066	-0.038	0.151	-0.118	-0.073	0.058
	(0.046)	(0.066)	(0.171)	(0.128)	(0.039)	(0.045)
Occ.: Precision/Armed force	0.144	0.051	0.281	-0.099	-0.103	0.094
	(0.035)	(0.052)	(0.141)	(0.093)	(0.033)	(0.039)
Log(liquid resources)		-0.156				
		(0.009)				
Musculoskeletal disab.			-0.374			
			(0.148)			
Respiratory disab.			-0.558			
			(0.203)			

(continued from previous page)						
Cardiov. disab.			-0.235			
			(0.160)			
Endocrine disab.			0.074			
			(0.199)			
Neurol. disab.			-0.576			
			(0.192)			
Mental disab.			-0.496			
			(0.179)			
Cancer disab.			-1.066			
			(0.249)			
Immune def. disab.			-0.580			
			(0.249)			
Dig. and Urin. disab.			-0.844			
			(0.238)			
Female vignette				-0.048		
				(0.019)		
Vign.: Low pain				0.000		
				(.)		
Vign.: Interm. pain				-1.984		
				(0.050)		
Vign.: Severe pain				-0.139		
				(0.032)		
Vign.: Low depression				-1.007		
				(0.036)		
Vign.: Interm. depression				-0.984		
				(0.036)		
Vign.: Severe depression				-2.000		
				(0.050)		
Vign.: Low cardiov				-0.766		
				(0.036)		
Vign.: Interm. cardiov				-0.696		
				(0.035)		
Log(income)				0.218		
				(0.037)		
Low cash-on-hand					0.031	0.112
					(0.008)	(0.009)
Govt. health ins.					0.019	-0.265
					(0.018)	(0.021)
Priv. health ins.					-0.170	0.020
					(0.022)	(0.026)
Spouse has health ins.					-0.124	0.536
					(0.020)	(0.024)
Full covg of doc. and drugs exp.						0.061
						(0.025)
Full covg of hosp etc. exp.						-0.982
						(0.018)
Has spouse with OOP> 0						-0.521
						(0.025)
Spouse's OOP						0.465
						(0.021)
Wave dummies	Y	Y	N	N	Y	Y
Year dummies	N	N	Y	N	N	N
Observations	86891	86891	10366	19143	51980	51980

G Robustness: Approximation of $\bar{S}(X, F)$

In this Appendix we report the results mentioned in footnote 30 in the main text.

Table G.2: Moments and Mechanisms Parameters

Moment	Linear	Quadratic	Cubic	Parameter	Linear	Quadratic	Cubic
β_F^L	0.041	0.041	0.041	π	-0.060	-0.060	-0.059
β_F^A	-0.106	-0.106	-0.106	γ	-0.111	-0.110	-0.108
β_F^V	-0.162	-0.162	-0.162	τ	0.267	0.267	0.260
β_F^C	0.187	0.187	0.187	θ	-0.313	-0.312	-0.306
$\beta_{F_V}^V$	-0.048	-0.048	-0.048	λ	0.321	0.320	0.313
σ_C	1.363	1.363	1.363	φ	0.206	0.206	0.205
$\rho_{L,A}$	0.493	0.493	0.493	σ_ω	0.732	0.722	0.683
$\rho_{A,R}$	-0.550	-0.546	-0.564	σ_φ	-0.171	-0.170	-0.167
$\rho_{L,R}$	-0.422	-0.428	-0.433	σ_v	1.625	1.626	1.569
$\rho_{L,V}$	0.0109	0.0109	0.0109	σ_ϕ	1.325	1.325	1.327
$\rho_{C,L}$	0.257	0.257	0.257	α_1	-0.626	-0.652	-0.644
$\rho_{C,A}$	0.155	0.155	0.155	σ_ξ	1.046	1.047	1.043
$\rho_{C,R}$	0.222	0.221	0.217				

Note: The left side of this table reproduces the results of Table 6 under three different assumptions for the approximation of the marginal signal $\bar{S}(X, F)$ in equation (8): a linear approximation around $(\bar{X}, F = 0)$ (the baseline), a quadratic approximation, and a cubic approximation. The right side of the Table reports the corresponding EWMD estimates of the mechanism parameters.