Outcome Tests and Screening Errors*

Hamish Low† Luigi Pistaferri‡

June 3, 2025

Abstract

This paper clarifies the conditions needed to test for discrimination, and shows how to test whether observed disparate treatment is taste-based or statistical. We use the context of applications for disability insurance where an authority (the SSA) decides whether to award benefits, possibly with a taste for discrimination against specific groups. We stress the need to specify the objective of the authority, to characterize their information set, and the importance of independent data on an individual's type. We focus on an authority with an objective to avoid Type I and Type II errors. First, we show that outcome tests à la Becker may fail to detect discrimination when the distribution of estimated eligibility differs for people who are truly eligible from those who are not – even in circumstances in which the distribution of "ground truth" is independent of group affiliation. Second, we show that self-reported data on eligibility (specifically, self-reported disability) and detailed data on the information set of the authority can distinguish between taste-based discrimination and statistical discrimination. We find strong evidence of taste-based discrimination against women in the award of disability benefits.

*This paper uses restricted HRS data made available to Pistaferri under confidential agreements RDA 2020-070. Thanks to our discussant Ivan Canay, and to Stephane Bonhomme, Magne Mogstad, and participants to the *Journal of Political Economy: Microeconomics Causal Inference Conference* for comments. All errors are ours.

†University of Oxford and Institute for Fiscal Studies.

‡Stanford University, SIEPR, NBER and CEPR.

1 Introduction

It is extremely difficult to detect taste-based discrimination in many settings: by employers, judges and evaluators, against people of a certain group. Taste-based discrimination means having different award thresholds in decision making that arise because of the preferences of the decision maker. The difficulty is because there are multiple reasons why award thresholds may differ. For example, individuals of different groups may be treated differently because of differences in other relevant characteristics that are correlated with group membership - statistical discrimination. Even when statistical discrimination can be ruled out or controlled for, taste-based discrimination may be hard to distinguish from inaccurate beliefs. The underlying aim of this paper is to provide a framework for analyzing discrimination and relating this framework to the data requirements for testing discrimination. We first discuss some limitations of current ways of disentangling taste-based discrimination. We then show how the econometrician can use information on an individual's type and the decision maker's information set to disentangle taste-based discrimination.

Our contribution is to focus on the errors created by an evaluator using limited or noisy information, and how understanding these errors can assist with tests of discrimination. Limited information leads to false rejections (Type I errors) and to false positives (Type II errors), and the extent of each depends on the thresholds (which may differ by gender or other characteristic). This perspective on the decision process highlights particular problems with using popular tests for detecting discrimination (such as outcome tests), and also suggests alternative ways of identifying discrimination.

To fix ideas, we use throughout the example of the award of disability insurance benefits by an evaluator in the Social Security Administration (see Low and Pistaferri (2025)). Awards are intended to be made to those with health conditions that limit work. Individuals differ by whether they actually have a work limitation covered by the program, but this status is only observed by the evaluator with noise. Awards are made to those whose perceived disability, given the information set, is above a threshold. The question is whether this threshold differs by gender and whether this difference arises because of taste-based discrimination or other reasons. The focus on gender is because rejection rates are higher for women than men.

A traditional way to measure discrimination or disparate treatment is the use of a "benchmark test".[1] In our example, a benchmark test compares rejection rates for men vs. women: if women are rejected at a higher rate than men (controlling for observable characteristics that do not generate an included variable bias), the benchmark test concludes that there is evidence of disparate treatment.[2]

[1] We will use the terms discrimination or disparate treatment interchangeably.

[2] An included variable bias may emerge if the researcher adds controls that are such close proxies for gender that

The advantage of a benchmark test is that it is based directly on the choices made by the evaluator who is suspected of discrimination. The main disadvantage of the test is omitted variable bias. If there are characteristics driving work limitations that are observed by the SSA but not by the researcher, and which are correlated with gender, then disparate treatment may reflect these differences in traits rather than explicit discrimination.

More generally, evaluators observe imperfect information on the true disability state of the applicant and form an expectation of disability conditional on the information set. The problem is that this expectation of being work limited may differ between men and women even in the presence of the same information set. This will lead to differences in thresholds for acceptance even without taste-based discrimination: this is statistical discrimination. Following the literature, we refer to the expectation of being work limited conditional on the information set (and gender) as the "posterior risk" of disability.

Becker (1971) proposed to solve the omitted variable bias problem with an "outcome test". The latter is based on actions consequent to the evaluator's decision - i.e., on outcomes related to the decision rather than on the decision itself. In our example, the outcome could be whether the rejected applicant returns to work or not. If there were no taste-based discrimination, the work rate of women who are rejected should be similar to that of men. However, if women were subject to a tougher rejection threshold than men, then we would find that the work rate would be lower for women than for men.

The problem with the outcome test in this example is that the distribution of the probability of working differs between men and women who have been denied benefits. If the distribution of the posterior risk of disability is different for men and women, average work rates of women may be lower than average work rates of men even if, at the margin, men and women are treated similarly. And *vice versa*, it is possible that the average work rates of men and women could be similar even if at the margin there is taste-based discrimination. This problem with the outcome test is known as the inframarginality problem: average differences do not imply differences at the margin.

There have been several approaches to resurrecting the use of the outcome test, addressing the inframarginality problem, and more generally, the problem of group differences in the posterior risk distribution. One is to consider an equilibrium setting in which both the demand and the supply are modeled and where average effects and marginal effects are the same (see Knowles et al., 2001). A second approach is to consider a marginal version of the outcome test (see Arnold et al., 2022). A different strategy is to estimate directly the thresholds using distributional assumptions (see

render the gender coefficient insignificant, i.e., a dummy for a history of breast cancer in our context. See Ayres (2005).

Simoiu et al., 2017). These approaches to the inframarginality problem focus on the issue being that the risk distribution differs by gender (or by the discriminating characteristic). None consider the possibility that the distribution of the posterior risk of disability may differ by the true state. Low and Pistaferri (2025) take account of this and propose an alternative approach: a conditional benchmark test, where the conditioning is on the true state and gives rise to separating out Type I and Type II errors.

This paper provides a framework of discrimination decision-making for disentangling this large literature, and makes two additional contributions.

Our first contribution is to argue that the outcome test may fail to reveal taste-based discrimination even in the absence of inframarginality (i.e. even if the distribution of work limitations is identical for men and women). This arises if the distribution of the posterior risk of disability for "false" applicants differs from the distribution of the posterior risk of disability for "truly disabled" applicants. To understand this, take the case where evaluators adopt higher standards for women than men and assume that the distribution of underlying health is identical for men and women. Type I errors will be larger for women than men but Type II errors will be smaller for women than men. The sample of rejected applicants will be a mixture of truly disabled people (with a larger fraction of women than men) and healthy people (also with a larger fraction of women). Depending on how much the two groups weigh in the sub-sample of rejected applicants, we may find that the average disability rate is lower (and hence the average employment rate higher) for women than men, which would be interpreted as absence of taste-based discrimination against women even in this case where it exists.

Our second contribution is to consider a different solution to the problem of identifying taste-based discrimination, which we call a conditional benchmark test, based on using self-reported work limitations data. First, we use the information set of the authority to predict the work limitation status of the applicant. The conditional benchmark test then infers any threshold differences by comparing the rejection rates of men and women controlling for this predicted work limitation status. Any role for gender (or other characteristic) over and above the predicted work limitation status is evidence of shifts in thresholds across groups.

In the example of applications for disability insurance the key issue is that the evaluators have only imperfect information about the work limitation status of the applicant. However, there are circumstances in which the true work limitation status may be observed by the econometrician even if it is unobserved (or observed with noise) by the evaluator. In our example, this information comes from survey data where people self-report the extent of work limitations they face. This information allows us to consider the posterior risk distribution for genuine claimants separately

from false claimants. The two key components of this approach are data on the true type and data on the information set of the evaluator. We argue that this situation has a wide range of applications, and provide various examples from the literature.

We start in Section 2 by outlining formally the decision problem of the evaluator. Section 3 shows how the outcome test can fail even if the distributions of posterior risk are identical for men and women. Section 4 shows how information on an individuals' underlying type can be used to test for taste-based bias. Section 5 concludes.

2 Modeling Disparate Treatment

In this section we outline how to model disparate treatment and the ingredients needed to disentangle taste-based from statistical discrimination in empirical applications.

Suppose that an authority z has to make a binary decision $D_i = \{0, 1\}$ that affects the welfare of an agent i. The decision is whether to provide the agent with a specific service they have requested or applied for. Eligibility (or qualification) for the service cannot be established with certainty by the authority because the state that determines eligibility, Y_i^*, is only imperfectly observed by the authority.[3] In Table 1 we offer several examples of this general setting. We focus mainly on the example of the award of disability insurance (DI) to illustrate the modeling of disparate treatment and the difficulty with establishing taste-based discrimination. Individuals who apply for DI are evaluated by officers in the Social Security Administration (SSA), z, for whether DI benefits are awarded or not. Awards are intended for those who have a health condition that leads to a work limitation (Y_i^*). The information set, $\mathcal{I}_i^z$, of the SSA contains measures of the applicant's health status, alongside demographic and labor market details. The SSA forms an expectation of the individual's type: $E(Y_i^*|\mathcal{I}_i^z)$. The award decision will depend on this expectation and, crucially, on the objective of the authority. Subsequent to the award decision, some individuals will return to work, and further their health may deteriorate or improve over time: these are outcomes, Q_i, of the award decision.[4]

The question of interest is whether z treats members of some groups differently than members of other groups (disparate treatment), and why. To keep things simple, we let G_i be a binary

[3]Eligibility for the service may also require some technical requirements, such as previous work history, but we assume these technical requirements are observed by the authority, the agent and the analyst, and so are omitted from now on.

[4]In Low and Pistaferri (2025) we consider a more general case in which work disability is a latent continuous variable, the signal depends on this latent variable, and the equivalent of the variable Y^* is an indicator for observing the latent work limitation variable crossing a threshold for disability insurance eligibility that would be chosen by a neutral evaluator.

indicator for a minority (or potentially discriminated) group in the population. In the example below $G_i = \{m, f\}$ is an indicator for female applicants.

What drives the decision of the authority? Why may the authority treat members of the $G = f$ group differently than members of the $G = m$ group? Economists try to rationalize disparate treatment by appealing to the economic model of discrimination, due to Becker (1957) (see also Arrow, 1973). Becker considered a framework in which a firm has to choose the number of workers to hire in a setting in which workers can be one of two groups, Blacks and whites, and the firm maximizes its utility, defined as the sum of profits and a disutility term from hiring minority workers.

In our context, we assume that the problem of the authority is to choose $D_i = \{0, 1\}$ for n applicants to maximize expected utility, which we assume has a positive weight on correct decisions and a negative weight on Type I and Type II errors:[5]

$$\begin{aligned} \max_{D_i=\{0,1\}} \quad & \sum_{i=1}^{n} E\left[u_{TP}\left(1-D_i\right)Y_i^* + u_{TN}D\left(1-Y_i^*\right)\right. \\ & \left. -c_1 D_i Y_i^* - c_2\left(1-D_i\right)\left(1-Y_i^*\right)|\mathcal{I}_i^z, G_i\right] \qquad (1) \\ s.t.\ B \quad \geq \quad & \sum_{i=1}^{n}\left(1-D_i\right)b(G_i) + k \end{aligned}$$

Note that the dependence of the right-hand side of equation (1) on $\mathcal{I}^z$ and G captures the potential dependence of each of the parameters u_{TP}, u_{TN}, c_1 and c_2 on $\mathcal{I}^z$ and G, as well as their role in forming expectations of Y_i^*. We separate out the dependence on G from dependence on the information set $\mathcal{I}^z$ for clarity, but the assumption is that G is in the full information set of z.

The term $\left(1-D_i\right)Y_i^*$ corresponds to awarding services to a truly eligible applicant (a truly positive case, which offers utility benefit u_{TP}), $D_i\left(1-Y_i^*\right)$ is denying the services to an ineligible individual (a truly negative case, which offers benefit u_{TN}). $D_i Y_i^*$ corresponds to denying a service to an individual who is eligible for it. This is a Type I error that has utility cost c_1. The term $\left(1-D_i\right)\left(1-Y_i^*\right)$ corresponds to awarding the service to an ineligible individual. This is a Type II error that has utility cost c_2. One of the benefits/costs could be normalized to 0, so that all other benefits/costs will be interpreted in relative terms.[6]

We assume that B is the total budget allocated to the DI evaluators which is devoted to paying

[5]Kleven and Kopczuk (2011) consider a setting in which the authority may also weight negatively the cost of eligible people not applying for benefits (a special form of Type I error). This may also vary with G if this "discouragement" effect is coming from perceptions of discrimination.

[6]The benefits of making a correct decision could in principle be normalized to zero (i.e., one could assume $u_{TP} = u_{TN} = 0$). However, this seems restrictive since an authority that weights more the "insurance" component of its decisions (granting the service to the deserving) will generally have $u_{TP} > u_{TN}$ (and vice versa for an authority that weights more the "moral hazard" component).

benefits b to awardees plus "discovery" costs k (i.e., the cost of consultative medical examinations, etc.).[7] In principle b could depend on G. For example, in the Social Security DI case, benefits paid to women could be lower than those paid to men because of shorter work histories and lower average lifetime earnings (AIME). This would lower the threshold for admissions of women (see below). In other settings this may work in reverse. For example, in the example of ER physicians deciding whether to order an expensive test that can detect (with almost certainty) whether a patient is suffering from a stroke, the cost of treating Blacks could be higher than the cost of treating whites because of differences in availability of medical insurance, and this may set a higher threshold for receiving an expensive diagnostic exam. Per se, this does not reflect taste-based discrimination (which is reflected in the values of the utility parameters) but a form of "institutional" group difference.

The expectations in the objective are defined using the information, $\mathcal{I}_i^z$, that the examiner has about applicant i. This information set contains information that may be informative about Y_i^*, as well as other information on the individual which is uninformative about Y_i^*. For the SSA, the information set (at this stage of the process) is only the information on the application form, including details of the applicant's health condition and past employment, as well as gender, race and marital status. These characteristics may affect the evaluator's prediction about the work limitations of the applicant as well as having a direct effect on the threshold set. An analyst with access to data on the application form will have the complete information set of the decision maker. Often however the analyst will have data only on a subset of the information set. The question is then whether the omitted variables from the information set impact the decisions concerning men and women differently, or whether the award thresholds for each are affected in similar ways.

The first order condition of the problem yields a threshold rule:[8]

$$\begin{aligned} D_i &= \mathbb{1}\left\{E\left(Y_i^*|\mathcal{I}_i^z, G_i\right) \leq \frac{u_{TN}+c_2+\lambda b}{u_{TN}+u_{TP}+c_1+c_2}\right\} \\ &= \mathbb{1}\left\{E\left(Y_i^*|\mathcal{I}_i^z, G_i\right) \leq \tau^*\left(\mathcal{I}_i^z, G_i\right)\right\} \end{aligned} \tag{2}$$

The term $E\left(Y_i^*|\mathcal{I}_i^z, G_i\right)$ is the posterior risk (of being work limited). The assumption is that $E\left(Y_i^*|\mathcal{I}_i^z, G_i\right)$ reflects an *accurate* posterior risk prediction by the authority – i.e., when forming the

[7] The problem of the DI evaluators could also be that of deciding how much to invest in k in order to improve the information set and reduce the variance of the noise (assuming, realistically, that the precision is an increasing, concave function of k). A clear example of this is the fact that SSA often asks applicants to submit to a consultative examination (which is paid for by SSA), or to fill in a special 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..

[8] To avoid clutter, we have suppressed the dependence of u_{TP}, u_{TN}, c_1 and c_2 on $\mathcal{I}^z$ and G, and the dependence of b on G.

expectation, the authority is using only their actual information set, not any distorted beliefs (due to stereotypes, etc.).[9]

The term $\tau^*\left(\mathcal{I}_i^z, G_i\right)$ is the threshold that is used by the evaluator to decide who to deny/award benefits. Anyone who is assessed to have a posterior risk of being work limited below $\tau^*\left(\mathcal{I}_i^z, G_i\right)$ is denied, and *vice versa*. In general, τ^* may depend on components of the information set $\mathcal{I}^z$. These components may be uninformative about Y_i^* or they may be components that impact τ over and above their role in forming the expectation $E(Y_i^*|\mathcal{I}_i^z, G_i)$.

When τ depends on components of the information set unobserved to the econometrician, there is unobserved heterogeneity on both sides of the inequality in the optimality condition (2). In this situation it is not possible to separate out the effect of gender on the threshold. This is the point in the discussion of Canay et al. (2024), where they label this case as the Generalized Roy model. This would arise, for example, if the SSA evaluator met applicants in person and their appearance then formed part of the evaluator's information set. This (unobserved to the econometrician) information may be orthogonal to forming $E(Y_i^*|\mathcal{I}_i^z, G_i)$ and yet be used by the evaluator to alter the threshold (reflecting additional biases). From now on, we will restrict attention to a simplified case where the threshold $\tau^*\left(\mathcal{I}_i^z, G_i\right) = \tau^*\left(G_i\right)$ depends only (and possibly) on gender. This is the Extended Roy model. The strength of this restriction depends crucially on the quality of the data the econometrician has on the information set of the evaluator.

Note that from equation (2), denials become more likely when the threshold increases. If the authority wants to maximize the provision of insurance it will assign higher weight to providing the service to eligible individuals (a higher u_{TP} and a higher c_1), resulting in a lower threshold and hence a more lenient attitude. In contrast, an authority that wants to minimize moral hazard will assign higher weight to avoiding provision to those who are ineligible (a higher u_{TN} and a higher c_2), resulting in a higher admission threshold and hence a stricter decision rule. Moreover, the authority may become stricter if the benefits to be paid to an average allowed applicant or the shadow price of the resources available for the program increase. Budget considerations are typically ignored in these models.

There is now a substantial literature that uses variation in judge leniency as an instrument for receipt of a benefit (e.g. Maestas et al., 2013, French and Song, 2014). Our framework provides an interpretation of this variation in leniency as being due to differences in the cost of Type I and Type II errors and in the benefit of making correct decisions. Judges may differ in attitudes towards incentive vs insurance trade-offs and this will generate heterogeneity in the award threshold.[10]

[9]Suppose instead that the evaluator has inaccurate priors. Then any taste-based bias may become hard to disentangle from inaccurate posterior risk (see Bohren et al., 2023).

[10]Modeling explicitly the role of this heterogeneity among judges requires judge identifiers.

2.1 Disparate Treatment

We use the framework above to show the causes of disparate treatment. Disparate treatment exists when different denial decisions are made for two individuals characterized by the same information set but different group affiliation. There are three cases of interest.

Case 1. Pure statistical discrimination, or discrimination based on posterior risk:

$$\begin{aligned} E\left(Y_i^*|\mathcal{I}_i^z, G_i = f\right) &\neq E\left(Y_i^*|\mathcal{I}_i^z, G_i = m\right) \\ \tau^*\left(G = f\right) &= \tau^*\left(G = m\right) = \tau^* \end{aligned}$$

In this case, individuals characterized by the same information sets (i.e., similar health indicators, occupation, etc.), apart from gender, will receive different denial decisions because the risk of being work limited is estimated to be different. Higher denial rates for women may happen if – in the population – women are less likely to suffer a disability than men, or if women have noisier disability signals. This would imply gender is informative about the likelihood of being a $Y_i^* = 1$ person.

Case 2. Pure taste-based discrimination, or discrimination based on thresholds:

$$\begin{aligned} E\left(Y_i^*|\mathcal{I}_i^z, G_i = f\right) &= E\left(Y_i^*|\mathcal{I}_i^z, G_i = m\right) = E\left(Y_i^*|\mathcal{I}_i^z\right) \\ \tau^*\left(G = f\right) &\neq \tau^*\left(G = m\right) \end{aligned}$$

In this case, the posterior risk is the same but admission thresholds differ by gender. From equation (2), this may happen if the components of utility differ by gender, or if average benefits do. In this case, although gender is part of the information set of the examiner, it is not informative about Y_i^* and yet is used in the decision rule.

Case 3. Both statistical discrimination and taste-based discrimination:

$$\begin{aligned} E\left(Y_i^*|\mathcal{I}_i^z, G_i = f\right) &\neq E\left(Y_i^*|\mathcal{I}_i^z, G_i = m\right) \\ \tau^*\left(G = f\right) &\neq \tau^*\left(G = m\right) \end{aligned}$$

The sign of the inequalities determines whether we observe disparate treatment in favor of men (against women) or *vice versa*.

In Figure 1 we offer a graphical depiction of the three cases of interest. The figure plots CDFs of the posterior risk $E(Y^*|\mathcal{I}_i^z, g)$ for different values of $E(Y^*|\mathcal{I}_i^z, g)$.[11] The fraction of the population

[11] For illustration, the posterior risk $E(Y^*|\mathcal{I}^z, g)$ is assumed to follow a Beta distribution (with appropriately chosen parameters).

of men and of women who are observed to be denied are given by the level of the horizontal lines labeled "Women's denial rate" and "Men's denial rate (scenario 1)".

The first case is when the CDFs of posterior risk differ by gender (the lines labeled $F\left(E\left(Y^*|\mathcal{I}^z, m\right)\right)$ and $F\left(E\left(Y^*|\mathcal{I}^z, f\right)\right)$), and so gender is informative about Y^*. The CDF lines intersect the observed denial rates at point C for women and D for men, giving the values on the x-axis of $E\left(Y^*|\mathcal{I}_i^z, f\right)$ and $E\left(Y^*|\mathcal{I}_i^z, m\right)$ for the marginal recipient, which corresponds to τ_f^* and τ_m^*. In this case, there is a common admission threshold, τ^* and this is a case of purely statistical discrimination case.

The second case is where the CDF of posterior risk is independent of gender (the line labeled $F\left(E\left(Y^*|\mathcal{I}^z, g\right)\right)$). This CDF intersects the observed denial rates at point A for men and point B for women. The different observed denial rates imply different admission thresholds (higher for women than men), shown by the vertical lines labeled "M's threshold" and "F's threshold". This is a pure taste-based discrimination case.

The final case is a combination of the two. To generate it, we assume that the CDFs of posterior risk differ by gender as in case 1, but men face now a lower denial rate (the horizontal line labeled "Men's denial rate (scenario 2)". The intersections are now at point E for men and point C for women, and this implies a higher admission threshold for women than men. This arises due to a combination of statistical and taste-based discrimination. Evaluated at the men's threshold, the implied denial rate that would be consistent with pure statistical discrimination would be point F. The fact that women's denial rate is higher is evidence for an additional impact of gender through a higher threshold and for the presence of taste-based discrimination.

2.2 Components to Disentangle Disparate Treatment

As outlined in the Introduction, the two main approaches to disentangling discrimination using this framework rely either on looking directly at the decisions taken by the authority, or by looking at outcomes following the decision. There are four key ingredients to this evaluation by an econometrician, which we outline here. In the following two main sections, we show the role of these components.

1. The objective of the decision maker, and how the observed outcome, Q_i, following the decision relates to this objective.

   To evaluate the presence of taste-based discrimination, we have to take a stand about the objective of the decision maker. We distinguish the objective of the decision maker who does not have preferences for discrimination and then show the cost to such a decision maker of allowing a biased decision maker to evaluate the applicant. The model outlined above provides

one natural characterization of the objective, but the point is more general: we cannot evaluate whether discrimination is in play without making an assumption about what the objective of the decision is.

2. The true type or condition of the individual, Y_i^*, and how and whether this is observed by the econometrician. The main issue is whether the measure of Y_i^* is truncated, and in particular only observed for those where $D_i = 0$ or only for those where $D_i = 1$, and whether or not this is observed contemporaneously with the time of the decision. A further issue is whether the measure of Y_i^* is discrete or continuous.

   With DI, the individual's true type can be thought of as the continuous work limitation status, which may nonetheless be observed only in discrete form. This information may be observed by the econometrician in survey data independently of the decision making authority. Alternatively, this information on true health may be observed subsequent to the DI decision in ongoing health records. Truncation can arise, for example, if Y_i^* were only observed for those on medicare, and medicare is only given to those with $D_i = 1$. More commonly, in the pre-trial release example, pre-trial misconduct can only be observed for those who have been released and so the observation of Y_i^* is truncated. There is an additional issue of timing: a correct evaluation of taste-based discrimination requires a measure of Y_i^* at the time of the decision partly because Y_i^* may evolve and partly because the award, D_i, itself may have an impact on Y_i^*.

3. The information set, $\mathcal{I}^z$, that the authority uses to form expectations about an applicant's true work limitation, and how much of this information set the econometrician observes. Knowing the information set allows the econometrician to infer the distribution of $E(Y_i^*|\mathcal{I}^z, G)$.

   For the example of DI, the information set on the applicant is whatever is contained in the application package, and there is no face-to-face or other interaction. This is exploited in Section **??** below and in Low and Pistaferri (2025). However, for the example of the pre-trial bail decision (see Arnold et al., 2022), where the judge observes the defendant in court, the judge may have additional information over and above the econometrician, such as the demeanor or physical appearance of the applicant. This information may be relevant to the judge's estimate of the posterior risk (or to their setting of the threshold τ^*).[12] For the econometrician, the absence of this information may lead to mischaracterizing the importance of statistical discrimination rather than taste-based discrimination.

[12] As mentioned above, if the unobserved information of the decision maker affects the threshold, τ, directly, then it is not possible to distinguish taste-based from statistical discrimination.

4. In performing an outcome test, the choice of the outcome(s) to be assessed in the post-evaluation period. The key issue here is whether the outcome test is based on Y_i^* (the unobservable state that enters the objective of the decision maker) or on Q_i, which may be correlated but not coincide with Y^*. Average and marginal outcome tests may give misleading results if the outcome being studied does not represent the objective of the authority.

   For the example of DI, a natural Q_i could be whether the individual is working or not after the application. Rejected applicants in poor health are unlikely to be working, compared to rejected applicants in good health: Q_i is therefore correlated with Y_i^*. However, work status is affected by multiple factors: for example, there may be gender discrimination by employers rather than the authority, and yet this would show up as taste-based discrimination using an outcome test. This is discussed further below.

3 Outcome Tests and Screening Errors

A traditional way to measure discrimination is to use a benchmark test, which simply compares denial rates for men and women, and argues that there is discrimination against women if their denial rate is higher than those of men (controlling for observable characteristics). In practice, the test is whether a female dummy is significant when running a regression of denials against observables and the female dummy. Becker (1993) argued that omitted variable bias could invalidate the benchmark test: differences in denial rates may be due to variables in the decision maker's information set that are unobserved by the econometrician but correlated with gender. In other words, the unobservable (to the econometrician) component in $\mathcal{I}^z$ may be correlated with gender. The question in thinking of the validity of the benchmark test is the extent that the observables in the econometrician's information set capture the information set of the examiner (as discussed in Section 2.2). We return to addressing this issue in Section 4.2.

Becker recommended the use of "outcome tests" as an alternative to looking at differences in decision rates. The idea is to look at differences in outcomes that are consequential to the decision. Suppose –using Becker's example – that bank loan officers are biased against Black applicants, in the sense that they only approve loans for Blacks with particularly strong creditworthiness, i.e., they use higher standards for Black vs. white applicants. This means that we expect the loan performance to be better for approved Black customers than for white customers.

The well-known problem with the outcome test is "inframarginality". It is possible for the average outcomes for the two groups to differ (and hence provide evidence in support of disparate treatment) even when the authority uses similar standards; or *vice versa*, to find that average

outcomes for two groups are similar (hence providing no evidence in support of disparate treatment) even when the authority uses different standards for the two groups. This is because the distribution of the posterior risk $E(Y_i^*|\mathcal{I}_i^z, G_i)$ may differ by gender, and this impacts the outcome variable, potentially providing misleading conclusions. See Simoiu et al. (2017) for an example. The inframarginality issue is what has led many authors to focus on "marginal outcome" tests (Arnold et al., 2022) or "threshold" tests (Simoiu et al., 2017), where the goal is to identify admission thresholds directly.

In this section we discuss three issues related to outcome tests. A first issue is that the average outcome test may fail to reveal taste-based discrimination even when there is no "inframarginality" by group (i.e., the distribution of posterior risk does not vary by gender), but there is "inframarginality" by "ground truth" (i.e., the distribution of posterior risk varies by the unobserved true disability itself).[13] The second and third issues apply to both average and marginal outcome tests: the difficulty of knowing which outcome should the outcome test be based on, and the difficulty of establishing taste-based discrimination by the authority when the outcome may be affected by taste-based discrimination by other agents.

3.1 The Potential Invalidity of the Outcome Test

Suppose there is taste-based discrimination against a certain group. As discussed in the literature, an average outcome test can fail to reveal this discrimination under inframarginality, i.e., if the distribution of the ground truth (Y_i^*) or of posterior risk ($E(Y_i^*|\mathcal{I}_i^z, G_i)$) differ by group. In this section we show that the failure of the outcome test to reveal taste-based discrimination exists even in circumstances in which the distribution of the ground truth (Y_i^*) or of posterior risk ($E(Y_i^*|\mathcal{I}_i^z, G_i)$) are *identical* for both groups, as long as the distribution of posterior risk varies with the true unobserved disability. The key idea is that if there is taste-based discrimination, this changes the gender composition of those who are denied, and may well provide misleading results. Surprisingly, most of the discussion about the invalidity of the outcome test (inframarginality) ignores this possibility.

When an SSA examiner makes a decision regarding whether to award SSDI/SSI benefits to an applicant, it can make two types of errors: Type I errors (turning down for benefits an applicant who is truly work limited) and Type II errors (awarding benefits to an applicant who is not truly work limited). Below we show that differences between groups in screening errors are intimately related to the failure of the outcome test to reveal taste-based discrimination.

[13] Canay et al. (2024) make a similar point in their criticism of outcome tests (see their equation (34)). We show how to interpret this failure of outcome tests as being due to the extent of screening errors.

To restate the decision problem of Section 2, we assume that an SSA examiner can either deny a disability insurance application or not $D = \{1, 0\}$. Applicants are male or female, $G = \{m, f\}$. The unobservable attribute is whether applicants are actually work limited: $Y^* = \{1, 0\}$. People whose application has been denied can either work or not in the year(s) following rejection ($W = \{1, 0\}$). Working is thus the basis of the "outcome" test among rejected people.[14]

Suppose first that work limitations are independent of gender:

$$\Pr(Y^* = 1|G = m) = \Pr(Y^* = 1|G = f) = \pi$$

This is an important restriction because it says that there are no inframarginality issues that should make the outcome test invalid (or require a marginal version of the outcome test). In fact, this restriction implies that the benchmark test itself would work because the unobserved eligibility, Y^*, does not depend on gender.

Note that the employment rate of rejected applicants of gender g is:

$$\begin{aligned}\Pr(W = 1|G = g, D = 1) &= \Pr(W = 1|G = g, D = 1, Y^* = 1)\Pr(Y^* = 1|G = g, D = 1) \\ &\quad + \Pr(W = 1|G = g, D = 1, Y^* = 0)(1 - \Pr(Y^* = 1|G = g, D = 1))\end{aligned}$$

To take an extreme (knife-edge) case, suppose that truly work disabled people who are rejected never work ($\Pr(W = 1|D = 1, Y^* = 1) = 0$), while rejected applicants who are not work-limited always work ($\Pr(W = 1|D = 1, Y^* = 0) = 1$), and assume that these events are independent of gender. Hence:[15]

$$\Pr(W = 1|G = g, D = 1) = 1 - \Pr(Y^* = 1|G = g, D = 1)$$

The rate at which applicants of gender $G = g$ work is then determined by the types of error made by the SSA examiner and the probability of not being work disabled (and hence being able to work):

$$\begin{aligned}\Pr(W = 1|G = g, D = 1) &= 1 - \Pr(Y^* = 1|G = g, D = 1) \\ &= 1 - \frac{\Pr(D = 1|Y^* = 1, G = g)\Pr(Y^* = 1|G = g)}{\Pr(D = 1|G = g)} \\ &= \frac{(1 - T_{II}^g)(1 - \pi)}{T_I^g \pi + (1 - T_{II}^g)(1 - \pi)} \qquad (3)\end{aligned}$$

[14] In reality, awarded people can also work and retain their benefits as long as they work below the SGA amount; we will not focus on this group, which in any case is rather small.

[15] The traditional way in which outcome tests are conducted is to look at differences in average Y^* between men and women, so this is not exactly a knife-edge case. This is because the "outcome" one looks at and the variables whose posterior risk is assessed are one and the same. The condition we provide below would be unchanged if we assumed that the "causal effect" of work disability on employment in the denied population does not vary with gender, i.e., $E(W|D = 1, Y^* = 1, g) - E(W|D = 1, Y^* = 0, g)$ is independent of g.

The second line comes from application of the Bayes' theorem, $T_{II}^g = \Pr(D = 0|Y^* = 0, G = g)$ is Type II error (awarding when not disabled), and $T_I^g = \Pr(D = 1|Y^* = 1, G = g)$ is the Type I error (denying when truly disabled). Taste-based discrimination means that thresholds differ because of preferences and this translates into Type I errors for women being higher than for men, and Type II errors for women being smaller than for men. We can recast the outcome test using Type I and Type II errors to show where the outcome test may not be reliable.

The outcome test is based on testing whether the share of rejected women who work differs significantly from, and in particular, is lower than, the share of rejected men who work. Hence, the outcome test will be invalid and the test will fail to find evidence for taste-based discrimination against women whenever rejected women work at least as much as rejected men, or (after some algebra) if:

$$\begin{aligned} \frac{1 - T_{II}^f}{T_I^f} &\geq \frac{1 - T_{II}^m}{T_I^m} \\ \frac{\Pr\left(E(Y^*|\mathcal{I}^z, f) \leq \tau_f^*|Y^* = 0\right)}{\Pr\left(E(Y^*|\mathcal{I}^z, f) \leq \tau_f^*|Y^* = 1\right)} &\geq \frac{\Pr\left(E(Y^*|\mathcal{I}^z, m) \leq \tau_m^*|Y^* = 0\right)}{\Pr\left(E(Y^*|\mathcal{I}^z, m) \leq \tau_m^*|Y^* = 1\right)} \end{aligned} \tag{4}$$

The LHS is the ratio of true negative to false negative women; the RHS is the ratio of true negative to false negative men. Application of the Bayes theorem and the assumption that the distribution of Y^* does not differ by gender imply that (4) simply requires that: $\Pr(Y^* = 0|D = 1, f) \geq \Pr(Y^* = 0|D = 1, m)$. In other words, the outcome test will fail if the share of rejected undeserved applicants ("true negatives") among women is larger than the share of true negatives among men. Online Appendix Figure OA.1 shows a simple hypothetical example where the outcome test is shown to fail using equation (4). In that example, the RHS of equation (4) is equal to 1 (since the areas of false negatives and true negatives for men are both equal to A), while the LHS is greater than 1 (since for women the area of false negatives is (A+B+C) and the area of true negatives is (A+B)).

Condition (4) is satisfied in many realistic cases. For example, assume that Y^* is a draw from a Bernoulli distribution with parameter $\pi = \Pr(Y^* = 1|g) = \Pr(Y^* = 1)$, independent of gender (so that, no inframarginality issues arise). Next, we generate $E(Y^*|\mathcal{I}^z, g)$ from a Beta distribution with parameters $(\alpha_0(1 - Y^*) + \alpha_1 Y^*, \beta_0(1 - Y^*) + \beta_1 Y^*)$, again independent of gender. We choose the parameters of the Beta distribution such that $E(Y^*|\mathcal{I}^z, g)$ is increasing in Y^* (as it should) and has higher variance for $Y^* = 1$. In particular, we assume: $\pi = 0.5, \alpha_0 = 3, \beta_0 = 5, \alpha_1 = 1.5, \beta_1 = 1.8, \tau_f^* = 0.5$ and $\tau_m^* = 0.41$.

In this case, one can show that we can obtain a violation of the outcome test for realistic Type I and Type II errors (indeed, we find $T_I^f = 0.57, T_I^m = 0.45, T_{II}^f = 0.23, T_{II}^m = 0.40$). For these

parameter values the share of rejected applicants who work is 57% for both men and women (even though rejection rates are 67% for women and 53% for men). A researcher would conclude that the outcome test provides no evidence that women are discriminated against even though they face higher admission standards ($\tau_f^* > \tau_m^*$). See Figure 2.

Note that, in this example, the distribution of the posterior risk of disability is more precise for undeserved applicants (those with $Y_i^* = 0$) than for those who are truly work limited. This is a condition potentially observed in real applications, since it may be easier for work limited applicants to document a disability they actually have than for non-disabled applicants to document a disability they do not have. Or the work limitation itself may reduce the ability of the applicant to produce evidence. The point is that the SSA would have noisier information about them being disabled. This heterogeneity in precision is not heteroskedasticity by gender, it is by Y_i^*.

The intuition for this result is that, with lower variance when $Y^* = 0$, more women from the $\{Y^* = 0, G = f\}$ group are rejected than men from the $\{Y^* = 0, G = m\}$ group if they face different thresholds. This implies that the pool of rejected people will be overwhelmingly made up of women, who will be more employed than men (or employed as much as men).[16]

Note that while we may find evidence against taste-based discrimination even when there is some, the reverse is not true. If $\tau_m^* = \tau_f^*$, then type I and type II errors will be the same for men and women because the posterior risk distributions are identical. The outcome test will find that men and women work at similar rates and no evidence of discrimination. A violation may occur if $E(Y_i^*|G = 1) \neq E(Y_i^*|G = 0)$, but this would be a traditional inframarginality issue.

3.2 What is the outcome in the outcome test?

Many outcome tests in the literature are based on the same variable that is the subject of discovery by the evaluator (or judge). For example, in the bail example of Arnold et al. (2022), the judge's release decision depends on forming expectations about Y^* (criminality or *potential* misconduct) and the outcome test looks at *actual* misconduct among the released ($Y = Y^*(1 - D)$). Note that this example highlights the issue of truncation of observations of Y^* that was raised in Section 2.2, and more generally to the truncation of observations of the outcome. One issue with this approach is that, at least in principle, there is no reason to base the outcome test on a single outcome. For example, in the disability insurance application example, the SSA evaluator needs to assess whether an individual is able "*to engage in any substantial gainful activity (SGA) because of a medically determinable physical or mental disability that is either expected to result in death or has lasted or*

[16]Failure of the outcome test stems from the fact that the distribution of the signal depends on the true disability state (besides the obvious mean shift). It follows that a version of the outcome test that conditions on the true disability status may be one way to address the issue. See Low and Pistaferri (2025) for an example.

is expected to last for a continuous period of at least 12 months". The decision to deny benefits is the "benchmark" and there are multiple outcomes one can in principle study, such as working and earning more than SGA, mortality, morbidity, subsequent self-reports of disability, etc. In the bail example, a narrow view about the judge's mandate is that they only evaluate if defendants are "flight risks" (i.e., whether they will show up at trial). In reality, a judge may take into account multiple outcomes, such as whether a released defendant is likely to affect the trial's proceedings with their actions (e.g., by tampering with evidence or witnesses), or to commit any additional crime during the pre-trial period if released. An even broader view is that the judge makes their decision based on an assessment of the defendant's guilt or innocence, and is more likely to detain a defendant if they believe the defendant to be guilty of the crime they are accused of. In this case, the outcome would be whether the defendant is eventually found guilty or not guilty of the crime. With multiple outcomes potentially available to study, more can be learned from combining the results of various outcome tests, while recognizing that some of these outcomes may be too far removed from the "discriminating" decision that is being studied. We leave these issues for future research.

3.3 Who is discriminating?

A different issue with the traditional use of the outcome test is that the link between the outcome studied and the original decision may be attenuated by actions taken by other agents who also influence the outcome. For example, even if an applicant is able to engage in work, she may not be observed working because of labor market frictions (a lack of job offers) or because of employer discrimination. Suppose SSA is not discriminating against women, but rejection rates are higher for them (for example because of statistical discrimination). The outcome test suggests that if a judge has a taste for bias, we should observe fewer women working. But suppose that the labor market discriminates against women, so fewer women work in the first place. Then we would conclude that the SSA evaluators are biased when they are not: there is employer discrimination, but not discrimination by the authority. One way to counter this would be to check if there is discrimination *before* going through the application process, since there may be no reason this would change because of a disability. See Low and Pistaferri (2025) for an example. Of course, there may be more discrimination by employers against disabled women than against disabled men, which invalidates this test. In the bail example, actual misconduct depends not only on being released, but also on being "caught", and this may potentially differ by race. For example – among those who are released on bail – Black defendants may be more likely to be monitored by police than white defendants. If arrest rates are higher among Blacks, the outcome test would suggest no

taste-based discrimination even when there is some.[17]

4 Identifying Taste Bias using Self-Reported Y^*: the Conditional Benchmark Test

This section discusses our second contribution: we show how to use information on Y^* to uncover taste-based discrimination while allowing for the presence of statistical discrimination. We use the example of identifying taste-based discrimination in disability insurance (SSDI and SSI) decisions using self-reported disability data. There are two related ways in which having data on Y^* can be used to determine the extent of taste-based bias. First, the information on Y^* can be used to estimate directly the posterior risk distribution, $E\left(Y^*|\mathcal{I}^z, G\right)$. This can then be used in the decision rule, equation (2), to pin down whether the differences in award rates are due to statistical discrimination alone (see Grossman et al., 2024, for a similar idea). Alternatively, one can use a semi-parametric approach to identify differences in admission thresholds directly by inverting the CDF of posterior risk at the observed gender-specific denial rates. The second benefit of having data on Y^* is that we can separate explicitly denials which are correct denials (i.e., of non-work limited applicants) from denials that are mistaken (i.e., of work limited applicants). This relates back to the definition of the objective of the evaluator in equation (1) defined over Type I and Type II errors. See Low and Pistaferri (2025).

We start with some background on the SSDI and SSI programs, present the data, and then turn to the specific identification strategies.

4.1 Background and Data

Background: The SSDI and SSI Programs

We examine gender disparities in disability insurance application outcomes. Two federal programs provide coverage against disability risk: the Social Security Disability Insurance (SSDI) and the Supplemental Security Income (SSI).

SSDI is an insurance program offering financial and healthcare support to eligible workers, their spouses, and dependents. It aims to protect against long-term health issues that significantly impair work ability. The assessment process evaluates an applicant's health status and remaining work capacity. Unlike Workers Compensation or private disability insurance, which cover temporary

[17]Even in the famous example of Becker on loan default, discrimination by employers, contractors, or neighbors could make it harder for Black households to be able to service a mortgage.

work-related issues, SSDI focuses on severe and persistent health impairments. However, accurately assessing health status and its impact on work ability presents challenges due to imperfect observability.

SSDI benefits are calculated similarly to Social Security retirement benefits. Eligibility criteria include: (1) Submitting an application, (2) Meeting work history requirements, (3) Observing a five-month waiting period out of the workforce, (4) Earning below the Substantial Gainful Amount (SGA) if working, and (5) Meeting medical requirements proving inability to work. The work history requirement inherently provides less coverage for individuals who primarily work from home, often women, though this gender disparity is not the focus of this study.

SSI is a safety net program for working-age individuals with disabilities who have limited income and resources. It uses the same disability definition as SSDI but has additional income and resource limits similar to the SNAP program.

Background: The Disability Determination Process

The disability determination process, shared by both SSDI and SSI applicants, follows a sequential evaluation. Applications are submitted to local Disability Determination Service (DDS) offices and assigned quasi-randomly to an adjudicative team comprising a medical or psychological consultant and a disability examiner. The evaluation process consists of four steps, divided into two main stages. At the Health Evaluation stage (steps 1-2), SSA determines if the applicant has a severe and persistent medical disability, defined as "inability to engage in substantial gainful activity (SGA) due to a medically determinable physical or mental impairment expected to result in death or last for at least 12 continuous months." If the condition matches a "listed impairment," benefits are awarded without further review. If not, the applicant's residual functional capacity is assessed. At the Economic Opportunity Evaluation stage (steps 3-4): SSA verifies if the individual can perform their past work. If not, it assesses if the applicant can perform any work suitable for their age, education, and skills.

Only about 35% of SSDI/SSI applicants are awarded benefits at the DDS stage. Rejected applicants can appeal, with approximately 56% doing so. The appeal process includes a reconsideration stage reviewed by a different DDS officer, with an 11% success rate. Further appeal options (Administrative Law Judges (ALJ) or higher hearing levels) have higher success rates but are not studied here due to data limitations at the DDS level. Further, the ALJ have a broader information set than the decision maker at the DDS stage because the ALJ see the applicants in person, rather than only having the information on the application form.

Data: The Health and Retirement Study (HRS)

The Health and Retirement Study (HRS) is a comprehensive panel dataset that focuses on household heads aged 50 and above. It is funded by the National Institute of Aging (grant n. U01AG009740) and conducted by the University of Michigan. Our analysis utilizes a harmonized version of the HRS compiled by the RAND Center for the Study of Aging, which includes biannual waves from 1992 to 2020. Key variables in the HRS dataset include: (a) Self-reported work limitations: This indicates the presence of an impairment or health problem that restricts the type or amount of paid work a respondent can perform. It also provides information on whether the condition is temporary and if it completely prevents work (see below for a more precise definition); (b) Specific health conditions: The dataset includes indicators for various health issues such as high blood pressure, diabetes, cancer, lung disease, heart disease, stroke, psychiatric problems, and arthritis. These conditions are reported as diagnosed by the respondent's physician. Additional health indicators are also available, such as difficulty with activities of daily living (ADLs); (c) Out-of-pocket medical expenses: This information is available from the third HRS wave onward. This rich dataset allows for a comprehensive analysis of health, work limitations, and associated factors among older adults in the United States.

Data: Social Security Administration Records

For approximately 80% of HRS respondents who have provided consent, their data can be linked to administrative records from the Social Security Administration. These records include: (a) Master Earnings File (MEF), (b) Master Beneficiary Record (MBR) file, (c) SSI beneficiaries file, and (d) Form 831 Disability Records (F831). The F831 database provides information on the initial medical determination of applications for SSDI and/or SSI, including the initial review and reconsideration at the SSA level. However, it does not include "technical denials" or decisions made at the Administrative Law Judge (ALJ) level or beyond. The F831 database contains multiple records per individual, organized into application cycles and rounds. Each cycle may include up to two rounds: the initial DDS assessment and, if applicable, the DDS reconsideration. Individuals may have several cycles over time.

For each application cycle, the following information is available: a) Exact application date for each round, b) Outcome and decision date for each round, c) Primary impairment (body system) code, d) Stage of denial or award, e) Type of application (SSDI, SSI, or concurrent SSDI/SSI), f) Additional information, such as whether a consultative examination was requested. This comprehensive dataset allows for detailed analysis of disability application processes and outcomes, linking them to individual health and demographic characteristics from the HRS. Given

that we observe the Social Security earnings history of each applicant, we can also compute expected SSDI benefits using the benefit formula and the application year.

4.2 The Measurement of Y^*

To perform the test we discuss below, we need a measure of an individual's "true" work disability status, i.e., a measure of the Y^* variable described above. We approximate the SSA's definition of work disability using three HRS survey questions:

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?"

3. "Does this limitation keep you from working altogether?"

We classify individuals as work disabled (or severely work limited) if they answer "Yes" to questions 1 and 3, and report that the condition is not temporary. This closely matches the SSA's criteria of work-related impairment, severity, and expected duration.

In Table 2 we present descriptive statistics for the sample of SSDI/SSI applicants, separately for men, women and the whole sample. On average, 54% of men are denied, as opposed to 65% among women, despite self-reported work limitation rates being higher among women. More men are applying with cardiovascular conditions, and there is correspondingly a higher rate of high blood pressure, heart condition, and stroke diagnoses. In contrast, women are more likely than men to be diagnosed with arthritis and lung disease. In terms of occupation, men are more likely to be observed in blue collar occupations, and women are more likely to be employed in clerical or service occupations. Men and women are asked to submit to a consultative examination (a post-application medical or psychological examination that the SSA requests when the applicant's medical records are outdated or incomplete) at roughly similar rates. Finally, due to longer attachment to the workforce and higher lifetime earnings, men have higher estimated SSDI benefits.

When is Y_i^* observed by the analyst?

Our analysis is predicated on the analyst observing Y_i^* even in circumstances in which the decision maker does not. In our example, the assumption is that self-reported work limitations in a survey like HRS are a measure of Y_i^*. The advantage of our setting is that the measure of Y_i^* is contemporaneous with the decision taken by the authority, z. More generally, it is often possible to measure Y_i^* subsequently to the decision. For example, consider the audit study conducted by Nagi (1969).

Following the DI application and decision process, a team of five experts (a doctor, a psychologist, an occupational therapist, a social worker, and a vocational counselor) visited a sample of 2,454 DI applicants, conducted extensive interviews and data gathering, and with the help of a moderator, reached a final decision regarding the disability status of the applicant, ignoring the actual SSA's award decision (and independently of it). One could think of this exhaustive investigation as providing a measure of Y_i^* (or perhaps as a much more precise signal of the true Y_i^*, which would be too costly and hence infeasible to conduct for each DI applicant).

Different examples come from other settings. Philip and Ozkaya (2025) consider ER physicians having to decide whether a patient is suffering from a stroke, the equivalent of Y_i^* above. A patient who is seeking a particular stroke treatment or diagnostic exam designed to ascertain the truth (such as a non-contrast computed tomography) must be assessed by the ER physician to have an in-progress ischemic or hemorrhagic stroke (i.e., the ER physician must form a posterior risk assessment). The ER doctor observes symptoms (such as limb weakness, facial drooping, speaking difficulty, headaches, dizziness, etc.), which all enter the information set $\mathcal{I}^z$ that the doctor uses to form $E(Y_i^*|\mathcal{I}_i^z)$. Whether a patient actually had a stroke can be inferred retrospectively (*via* neuroimaging) even if initially misdiagnosed, and this constitutes the true Y_i^*. In the medical literature there are many examples in which Y_i^* can be inferred ex-post (the most obvious case is through an autopsy).

An example in a different context is when a pre-trial judge makes a decision whether to allow bail for a defendant in a rape or murder case based on an objective concerning innocence of the defendant, where the judge may not want to incarcerate until trial someone who is innocent. This would give a notion of Y_i^* that is broader than the narrower objective of assessing the defendant's potential for misconduct between release and trial. If DNA evidence can be collected at a later stage, it could constitute a true measure of Y_i^* that can be used to exculpate the defendant or assess his guilt with near certainty. Of course, this later measure of Y_i^* may be affected by how vigorously the prosecutor and police pursue conviction, which may itself be subject to bias.

The reliability of information on Y_i^*

How informative/reliable are data on Y^* in our context? In the medical examples we provided above, there is usually a high accuracy ratio, since the truth can be ascertained ex-post with almost certainty.

Self-reported work limitation measures offer less objectivity. Indeed, the literature has argued that they have both advantages and drawbacks. Benitez-Silva et al. (2004) argue that these measures allow individuals to provide a comprehensive summary of their health and disabilities. However,

Bound and Burkhauser (1999) identify three main concerns: endogeneity with respect to labor market outcomes, interpersonal comparability issues, and perception errors. In our previous work (Low and Pistaferri, 2025) we use disability vignette responses to control for individual-specific heterogeneity in disability reports and verified our results using objective health indicators, as well as incorporating perception errors in our statistical model. The results we obtained were qualitatively similar.

In the Online Appendix (Table OA.1) we examine the association between self-reported work disability and several objective health indicators (Difficulties with Activities of Daily Living (ADLs), hospital stays, BMI, mortality, Doctor-diagnosed conditions, and Out-of-pocket health spending). Invariably, people who self-report a severe disability are more likely to also report more objective health conditions (including higher mortality rate) and higher out-of-pocket health expenditures. The patterns are consistent for both men and women and are robust to controlling for age.

4.3 Identification of Award Thresholds

In this section we show when it is possible to identify taste-based discrimination from statistical discrimination combining direct evidence on "ground truth" (Y_i^*) with data on the information set of the decision maker.

Our administrative SSA data provide information on the first round evaluator's decision as well as reconsiderations (appeals). We call $D = \{0, 1\}$ the SSDI/SSI denial indicator. From HRS survey data we observe $Y^* = \{0, 1\}$, an indicator of the true disability/work limitation status of an individual. It is important to note that Y^* is observed for all applicants (awarded and rejected), and this is crucial for the test below to identify taste-based discrimination. In the literature that detects discrimination using average or marginal outcome tests, researchers observe a truncated distribution of Y^* (i.e., in the bail example, misconduct by defendants is observed only for those who are released; in Becker's example, default rates by households are only observed for those who were granted a mortgage; etc.).

The analyst observes a vector of application information X_i. We assume that this approximates the information set $\mathcal{I}_i^z$ of the SSA decision maker. The assumption that we observe the information set is analogous to an assumption of selection on observables. While this may be a strong assumption in some settings, our dataset is unusually rich. First, we observe administrative information on the application process and on the earnings and employment history of each applicant; second, health data from the HRS supplement the part of the application process where the administrative data may be incomplete. For example, applicants assessed at the vocational stage of the DI determination process 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. We use the rich HRS data about specific difficulties with activities of daily living (ADL), doctor-diagnosed illnesses, etc., to reproduce the extra information that applicants provide to SSA.

We obtain an estimate of the posterior risk assessment, $E\left(Y^{*}|\mathcal{I}^{z},G\right)$, as the predicted value of a probit regression of Y^{*} on X and G. The posterior risk $E(Y^{*}|\mathcal{I}^{z},G)$ formed by the authority may be estimated with error by the analyst. This error would partly reflect the discrepancy between the information set of the authority and the set of variables we condition on, and partly functional form assumptions.

The probit regression is run separately by gender.[18] We plot the distribution of the posterior risk estimate separately for men and women in the Online Appendix, Figure OA.2. The regression results are reported in Table 3 (probit and logit estimates are almost identical). We find that, in the group of applicants, ADL variables, occupation variables, and body system variables are jointly significant. Blacks are less likely to report a work limitation. Aggregate variables also matter: in years in which the unemployment rate is higher or the DI application rate is lower, more applicants report to be work limited.

Armed with the estimate of posterior risk, we follow two approaches to identify award thresholds. The first is a semi-parametric approach. Using the objective outlined in Section 2, the decision rule of the evaluator is given by equation (2), reproduced here:

$$D_{i}=\mathbb{1}\left\{E\left(Y_{i}^{*}|\mathcal{I}_{i}^{z},G_{i}\right)\leq\tau^{*}\left(G_{i}\right)\right\}$$

and therefore, omitting the i subscript:

$$\Pr\left(D=1|G\right)=F_{E(Y^{*}|\mathcal{I}^{z},G)}\left(\tau^{*}\left(G\right)\right) \tag{5}$$

where $F_{E(Y^{*}|\mathcal{I}^{z},G)}(.)$ is the CDF of $E\left(Y^{*}|\mathcal{I}^{z},G\right)$.

In the data we observe the share of female and male applicants who are rejected, i.e.,

$$\Pr\left(D=1|G=g\right)=p_{g}$$

Combining with equation (5) but using the predicted CDF of the posterior distribution (i.e., substituting X for $\mathcal{I}^{z}$):

$$\begin{aligned} F_{E(Y^{*}|X,g)}\left(\tau^{*}\left(g\right)\right) &= p_{g} \\ \tau^{*}\left(g\right) &= F_{E(Y^{*}|X,g)}^{-1}\left(p_{g}\right) \end{aligned}$$

[18] Results obtained pooling data for both genders are qualitatively similar.

The crossing of the observed denial rate for gender g and the CDF of $E\left(Y^*|X,g\right)$ pins down the admission thresholds. This procedure does not require having decision data on individual judges, but it does require having data on self-reported disability by applicants. It identifies the threshold for gender g as the p_g-th quantile of the distribution of posterior risk (see, in a different context, Frandsen, 2015). In the next section we show how to decompose total bias (differences in denial rates) into taste-based discrimination and statistical discrimination using a simple Oaxaca-Blinder decomposition.

Figure 3 plots the CDF of the estimated $E(Y^*|X,G)$; the horizontal lines are the empirical denial rates for men and women. The difference in denial rates is 11 percentage points (0.65 vs 0.54, respectively for women and men). The points in which the empirical denial rates cross the CDF of the estimated $E(Y^*|X,G)$ pin down the gender specific thresholds τ_f^* and τ_m^*. Clearly, $\tau_f^* > \tau_m^*$. The estimated difference in thresholds $(\tau_f^* - \tau_m^*)$ is 0.12 and significant at any standard conventional level (a block bootstrap s.e. of 3 p.p.).[19]

A second approach for the identification of award thresholds is parametric. It is based on a (probit) regression for the decision to deny benefits. From equation (2), if gender differences in denial rates were merely a reflection of statistical discrimination, we should find no role for gender once we control for the estimated posterior risk.

The results (marginal effects) are reported in Table 4. Since the posterior risk is a generated regressor, we compute standard errors with the block bootstrap (which also account for serial correlation due to multiple applications by the same individual). Column (1) shows that a higher posterior risk of being work limited reduces the probability of being denied benefits, as expected. However, gender is still statistically significant even after controlling for $E(Y^*|X,G)$. Denial rates are 12 p.p. higher for women controlling for the posterior risk of disability.[20] In column (2) we check whether budgetary issues explain denials, and we do so by adding estimated DI benefits.[21] In fact, people who expect to receive higher benefits are, ceteris paribus, *less* likely to be denied, over and above the predictive power of SSA earnings history on the $E(Y^*|X,g)$.[22] Finally, recognizing that the utility benefits and costs of particular decisions may depend on variables other than gender, in column (3) we include other demographic controls that could (at least in principle) cause bias and

[19] We use the block bootstrap to account for the fact that some individuals submit multiple applications over our sample period.

[20] As discussed in Section 2, this interpretation relies on the assumption that the data are generated by an Extended Roy model. If instead there are components of the information set that are unobserved to the econometrician and that affect τ directly then it is not possible to empirically rule out the alternative Generalized Roy model (see Canay et al., 2024) and so we cannot identify the effect of gender on τ.

[21] These are calculated using an applicant's SSA earnings history and the AIME/PIA formulae for SSDI applicants, and maximum SSI benefits for SSI applicants.

[22] We obtain similar results if we drop SSI applicants.

shift the thresholds: race, marital status, welfare claimant status. Race and marital status do not seem to matter. However, evaluators appear to set lower standards for welfare claimants.

4.4 A Oaxaca-Blinder Decomposition

In Figure 3 we plot the CDF $F_{E(Y^*|X,G=g)}(x) = F_g(x)$ separately by gender. The difference in denial rates is $p_f - p_m = F_f(\tau_f^*) - F_m(\tau_m^*)$. This difference can be decomposed into two parts using a simple Oaxaca-Blinder style decomposition: (a) a taste-based discrimination term $(F_m\left(\tau_f^*\right) - F_m\left(\tau_m^*\right))$, where we keep the posterior risk distribution fixed (at the male values) but change the thresholds; and (b) a statistical discrimination term $(F_f\left(\tau_f^*\right) - F_m\left(\tau_f^*\right))$, where we keep the threshold fixed (at the women's value) but change the distribution of posterior risk.[23]

In Table 5, column (1) we show the decomposition (and also report, in the first row, the estimated difference in award thresholds for completeness). The total difference in denial rates (second row) is 11 percentage points higher for women than men. The remaining rows report the decomposition. In the third and fourth rows, we keep the posterior risk distribution fixed at the male values. Under this counterfactual scenario, there would be statistical discrimination *against* men (because they tend to have demographic characteristics and health conditions that, on average, produce higher denial rates). This means that taste-based discrimination against women is even higher than the total award rate difference (16 p.p.). As in a standard Oaxaca-Blinder decomposition, the decomposition is not invariant to which risk distribution or threshold we keep fixed. For example, a different (also valid) decomposition is: $p_f - p_m = \left[F_f\left(\tau_f^*\right) - F_f\left(\tau_m^*\right)\right] + \left[F_f\left(\tau_m^*\right) - F_m\left(\tau_m^*\right)\right]$. The first term is taste-based discrimination and the second is statistical discrimination. In the fifth and sixth rows we thus keep the posterior risk distribution fixed at the female values. The results are qualitatively similar, although both terms of the decomposition are now larger in absolute value.

4.5 Evidence at Reconsideration

The SSDI/SSI application outcome can be appealed. The first level of appeal is a "reconsideration", in which a different DDS examiner is assigned to the case and issues a decision. There are additional appeal levels beyond reconsideration, but we do not have data on those. Instead, we extend our analysis by looking at second round application outcomes.

Figure 4 is the reconsideration-level equivalent of Figure 3. It is constructed by obtaining new estimates of posterior risk $E(Y^*|\mathcal{I}, G)$ (since at the point of appeal both Y^* and the information set

[23] Of course, even if $p_f - p_m > 0$, it is possible for one of the two terms of the decomposition to be negative. For example, if $\left[F_f\left(\tau_f^*\right) - F_m\left(\tau_f^*\right)\right] < 0$ it means that, absent taste-based discrimination, statistical discrimination would produce *lower* denial rates for women than men.

of the examiner may be different from the initial stage).[24] At the reconsideration level applications are less successful. Denial rates are 83% for women and 77% for men. Given the empirical CDFs of posterior risk, we again find that admission thresholds are higher for women. However, at this stage the difference is smaller (5 p.p.) and statistically imprecise (a s.e. of 7 p.p.).

The Oaxaca decomposition (see Table 5, column (2)) shows that, at reconsideration level, there is much less (if any) statistical discrimination, and hence gender differences in denial rates can be entirely attributed to taste bias, although, as stressed above, they do not seem to be statistically significant. One possibility is that "errors" are corrected at the appeal level.[25]

5 Conclusions

Detecting taste-based discrimination à la Becker against specific groups is complex and a vast literature has evolved to tackle this problem both theoretically and with increasingly sophisticated econometric strategies. In most settings of interest, an authority has to decide whether to approve a service to an agent with identifiable demographic characteristics and imperfectly observed eligibility. Tests of discrimination are based either on comparing success rates ("benchmark tests") or post-decision outcomes ("outcome tests").

Our aim in this paper was to clarify and simplify the analysis of whether and when discrimination can be tested for, and whether observed discrimination represents threshold-based or statistical discrimination. We outline how the validity and feasibility of these tests depends on four key components: (1) specifying the objective of the authority; (2) data on the true type of the applicant; (3) the information set of the authority and how much of this information set is observed by the econometrician; (4) observations on the consequences of the decision. We stress in particular the role played by the authority having an objective that aims to avoid making Type I and Type II errors.

The context we study is the decision to award disability insurance benefits to potentially work-limited applicants. Award rates are lower for women than men, which begs the question of whether and why there is disparate treatment by gender. We show that data on self-reported disability (or, more generally, data on "ground truth" collected ex-ante or ex-post) is useful in two ways: they can be used to obtain a measure of the extent of screening errors (Type I and Type II errors) in the decision-making process of the authority, and they can be used to estimate posterior risks of disability.

[24] Due to reduced sample size and confidentiality issues we run the probit regression on a pooled sample.

[25] This result is confirmed when we run a probit regression for denial at reconsideration against the posterior risk estimate and the female dummy. The latter has a marginal effect estimate of 0.06 with a s.e. of 0.04.

We make two contributions. First, we show that traditional outcome tests can fail to reveal taste-based discrimination –even when it exists and even if true eligibility is independent of group characteristics–if the distribution of posterior risk differ by true work-limitations. The presence of different award thresholds changes the composition of "rejected applicants" in terms of work limitations: a tougher threshold means that not only will there be more work limited applicants rejected, but also more false claimants. These can balance out and the outcome test can show no evidence of discrimination.

Second, we show the importance of considering the information set of the authority. In situations where the information set is observed well by the analyst, we can estimate the posterior risk that an optimizing authority uses. We can use these predictions alongside information on denial rates to pin down directly the award thresholds and whether these vary by gender. In our application to disability benefit awards, similarly to Low and Pistaferri (2025), we find significant evidence that admission thresholds differ by gender, providing evidence of bias against women over and above the impact of gender through statistical discrimination (which, if anything, would produce *lower* denial rates for women).

Our approach of defining and estimating the information set of the authorities, and separating out estimates of the posterior risk distribution from estimates of the impact on thresholds, has many other applications, particularly in medicine, but which we leave to future research and researchers.

References

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

Arrow, K. (1973). The theory of discrimination. In O. Ashenfelter and A. Rees (Eds.), *Discrimination in Labor Markets*. Princeton: Princeton University Press.

Ayres, I. (2005). Three tests for measuring unjustified disparate impacts in organ transplantation: The problem of" included variable" bias. *Perspectives in biology and medicine 48*(1), 68–87.

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

Becker, G. S. (1957). *The Economics of Discrimination*. Chicago University Press.

Becker, G. S. (1993). Nobel lecture: The economic way of looking at behavior. *Journal of Political Economy 101*(3), 385–409.

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.

Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2023, 09). Inaccurate statistical discrimination: An identification problem. *The Review of Economics and Statistics*, 1–45.

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.

Canay, I. A., M. Mogstad, and J. Mountjoy (2024). On the use of outcome tests for detecting bias in decision making. *Review of Economic Studies 91*(4), 2135–2167.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2020). Are Referees and Editors in Economics Gender Neutral? *Quarterly Journal of Economics 135*(1), 269–327.

Frandsen, B. R. (2015). Treatment effects with censoring and endogeneity. *Journal of the American Statistical Association 110*(512), 1745–1752.

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

Grossman, J., J. Nyarko, and S. Goel (2024). Reconciling legal and empirical conceptions of disparate impact: An analysis of police stops across california. *Journal of Law & Empirical Analysis 1*(1), 118–133.

Kleven, H. J. and W. Kopczuk (2011). Transfer program complexity and the take-up of social benefits. *American Economic Journal: Economic Policy 3*(1), 54–90.

Knowles, J., N. Persico, and P. Todd (2001). Racial bias in motor vehicle searches: Theory and

evidence. *Journal of Political Economy 109*(1), 203–229.

Low, H. and L. Pistaferri (2025). Disability insurance: Error rates and gender differences. *Journal of Political Economy*. Forthcoming.

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.

Morrison, E. R., B. Pang, and A. Uettwiller (2020). Race and bankruptcy: Explaining racial disparities in consumer bankruptcy. *The Journal of Law and Economics 63*(2), 269–295.

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

Philip, M. and O. Ozkaya (2025). Disparate treatment and outcomes in emergency departments: Evidence from Florida. Unpublished manuscript.

Simoiu, C., S. Corbett-Davies, and S. Goel (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics 11*(3), 1193–1216.

Table 1: Situations for Evaluating Discrimination

Authority (z)	**Agent** (i)	**Service** (D_i)	**Unknown state** (Y_i^*)	**Outcome** (Q_i)
SSA examiner	DI applicant	DI benefits	Work limitation	Work Status
Pretrial judige	Defendant	Pretrial release	Criminality	Pre-trial misconduct
Bankruptcy judge	Debtor	Chapter 7/13 discharge	Good faith filer	Re-filing
Bank loan officer	Borrower	Mortgage	Creditworthiness	Default
ED Physician	Patient	Exam prescription	Stroke in progress	Re-admission
Editor	Author	Paper publication	Paper quality	Citations
FDA inspector	Pharma company	Novel drug approval	Drug safety/efficacy	Adverse effect reports
Health inspector	Restaurant owner	Restaurant rating	Food safety	Food poisoning reports

Note: For examples, see Low and Pistaferri (2025) (DI benefits), Arnold et al. (2022) (Pretrial release), Morrison et al. (2020) (Personal bankruptcy), Becker (1993) (Mortgage loans), Philip and Ozkaya (2025) (Stroke detection), Card et al. (2020) (Paper submission).

Table 2: Descriptive statistics

	Men		Women		All	
	Mean	SD	Mean	SD	Mean	SD
Denied benefits	0.54	0.50	0.65	0.48	0.61	0.49
Y^*	0.44	0.50	0.51	0.50	0.48	0.50
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
Lab. mark. experience	32.89	9.90	26.44	10.65	29.04	10.83
Applied SSI only	0.24	0.43	0.27	0.45	0.26	0.44
Applied DI and SSI	0.20	0.40	0.22	0.41	0.21	0.41
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
Age	57.14	4.39	55.84	5.93	56.36	5.40
Body System Code:						
Respiratory	0.05	0.21	0.07	0.26	0.06	0.24
Cardiovascular	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
Neurological	0.08	0.27	0.07	0.26	0.08	0.27
Mental disorder	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 deficiency	0.03	0.16	0.02	0.15	0.03	0.16
Digestive and Urinary	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
Doctor diagnosed:						
High blood pressure	0.06	0.23	0.03	0.17	0.04	0.20
Psychological condition	0.04	0.19	0.06	0.23	0.05	0.21
Heart condition	0.07	0.25	0.02	0.13	0.04	0.19
Arthritis	0.25	0.43	0.31	0.46	0.28	0.45
Diabetes	0.20	0.40	0.17	0.37	0.18	0.39
Lung disease	0.10	0.30	0.17	0.38	0.14	0.35
Stroke	0.11	0.31	0.07	0.25	0.08	0.28
Cancer	0.10	0.30	0.12	0.32	0.11	0.31
Difficulty w/ ADL:						
Walking	0.13	0.34	0.18	0.38	0.16	0.37
Dressing	0.26	0.44	0.26	0.44	0.26	0.44
Stooping, kneeling, crouching	0.68	0.47	0.77	0.42	0.73	0.44
Getting out of bed	0.21	0.41	0.27	0.44	0.25	0.43
Spent some nights in hosp.	0.52	0.50	0.44	0.50	0.47	0.50
BMI	29.21	6.17	30.86	7.53	30.19	7.06
Occupation:						
Unknown/missing	0.11	0.31	0.11	0.32	0.11	0.31
Managerial	0.06	0.24	0.05	0.22	0.06	0.23
Professional	0.07	0.25	0.10	0.30	0.08	0.28
Sales	0.05	0.23	0.07	0.26	0.06	0.25
Clerical	0.04	0.21	0.17	0.38	0.12	0.33
Protection services	0.03	0.17	0.02	0.13	0.02	0.15
Food prep. services	0.03	0.17	0.07	0.25	0.05	0.22
Health services	0.01	0.09	0.08	0.28	0.05	0.22
Personal services	0.07	0.26	0.14	0.35	0.12	0.32
Farming/forestry/fishing	0.03	0.18	0.02	0.12	0.02	0.15
Mechanics/repair	0.05	0.23	0.00	0.05	0.02	0.15
Construction/Mining	0.12	0.32	0.00	0.05	0.05	0.22
Precision craft	0.07	0.25	0.04	0.20	0.05	0.22
Machine operator	0.07	0.25	0.08	0.27	0.07	0.26
Transport operator	0.15	0.36	0.03	0.16	0.08	0.27
Handlers, etc	0.03	0.18	0.02	0.12	0.02	0.15
Requested consultative exam	0.47	0.50	0.50	0.50	0.49	0.50
Annual OOP medical spending	3370	6380	3410	7280	3400	6930
Estimated monthly DI benefits	11437	5674	8441	3421	9645	4699
Observations	645		960		1605	

Table 3: Regressions for Y^*

	Probit		Logit	
	Women	Men	Women	Men
College degree	-0.061	0.020	-0.064	0.022
	(0.040)	(0.042)	(0.041)	(0.042)
Black	-0.053	-0.108**	-0.054	-0.107**
	(0.037)	(0.043)	(0.036)	(0.043)
Lab. mark. experience	-0.003	0.001	-0.003	0.001
	(0.002)	(0.003)	(0.002)	(0.003)
Applied SSI only	-0.012	-0.075	-0.013	-0.072
	(0.040)	(0.051)	(0.040)	(0.052)
Applied DI and SSI	-0.054	-0.031	-0.055	-0.032
	(0.043)	(0.051)	(0.043)	(0.052)
Married	0.038	-0.071	0.039	-0.068
	(0.036)	(0.044)	(0.036)	(0.045)
Widowed	0.025	-0.023	0.026	-0.014
	(0.058)	(0.094)	(0.058)	(0.093)
Age	0.001	-0.005	0.001	-0.004
	(0.003)	(0.005)	(0.003)	(0.005)
Requested cons. ex.	0.022	0.016	0.023	0.018
	(0.034)	(0.038)	(0.034)	(0.038)
OOP medical spend.	0.003	0.007**	0.003	0.007*
	(0.002)	(0.003)	(0.002)	(0.003)
Appl. rate	-0.124*	-0.287***	-0.124*	-0.289***
	(0.067)	(0.079)	(0.067)	(0.079)
Unempl. rate	-0.003	0.042***	-0.003	0.043***
	(0.012)	(0.014)	(0.012)	(0.014)
Occupation (p-value)	0.000	0.008	0.000	0.003
Body system code (p-value)	0.114	0.471	0.110	0.345
Doctor diagnoses (p-value)	0.446	0.678	0.374	0.726
ADL (p-value)	0.000	0.000	0.000	0.000
Observations	960	645	960	645

Note: 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 and Personal services; 8. Farming, forestry, fishing; 9. Mechanics and repairs, Construction, trade and extractors, and Precision production; 10. Operators: machine; 11. Operators: transport, etc.; 12. Operators: handlers, etc.; 13. Others/Unknown. The Body system 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.

Table 4: Determinants of Denials

	(1)	(2)	(3)
$E(Y^*\|\mathcal{I}, G)$	-0.222***	-0.192**	-0.194**
	(0.071)	(0.071)	(0.075)
Female	0.126***	0.090***	0.082***
	(0.028)	(0.030)	(0.030)
Estimated DI benef.		-0.133***	-0.187***
		(0.033)	(0.037)
Black			0.030
			(0.032)
Married			0.006
			(0.027)
Applied SSI only			-0.133***
			(0.029)
Observations	1605	1605	1605

Table 5: Oaxaca decomposition

	First round	Reconsideration
Difference in estimated thresholds $(\tau_f^* - \tau_m^*)$	0.12***	0.05
	(0.04)	(0.07)
Total difference in rejection rates $(F_f(\tau_f^*) - F_m(\tau_m^*))$	0.11***	0.06
	(0.03)	(0.04)
Taste-based discrimination $(F_m(\tau_f^*) - F_m(\tau_m^*))$	0.16***	0.06
	(0.04)	(0.07)
Statistical discrimination $(F_f(\tau_f^*) - F_m(\tau_f^*))$	-0.04	-0.00
	(0.04)	(0.06)
Taste-based discrimination $(F_f(\tau_f^*) - F_f(\tau_m^*))$	0.25***	0.09
	(0.07)	(0.10)
Statistical discrimination $(F_f(\tau_m^*) - F_m(\tau_m^*))$	-0.14***	-0.03
	(0.06)	(0.10)

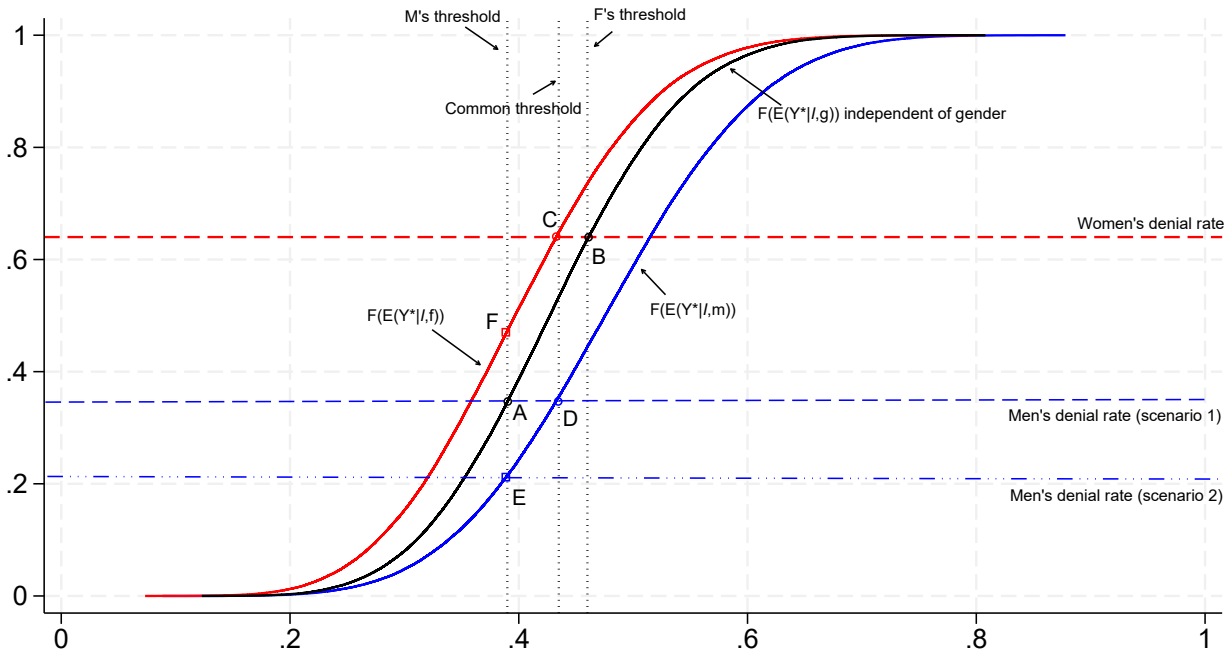

Figure 1: Three cases of interest.

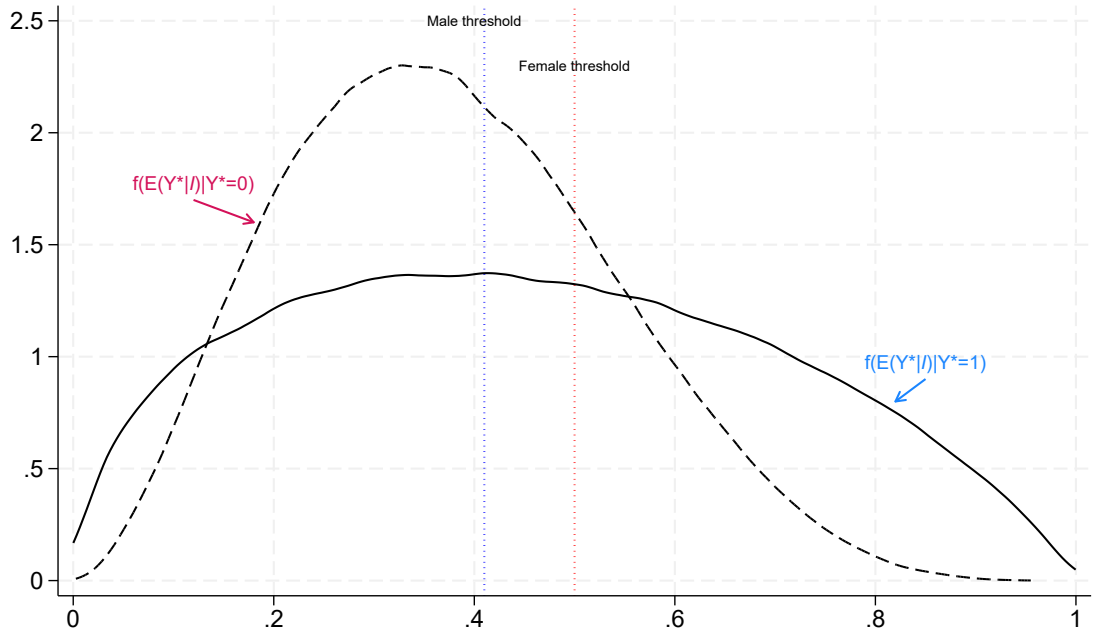

Figure 2: An example where the outcome test would fail.

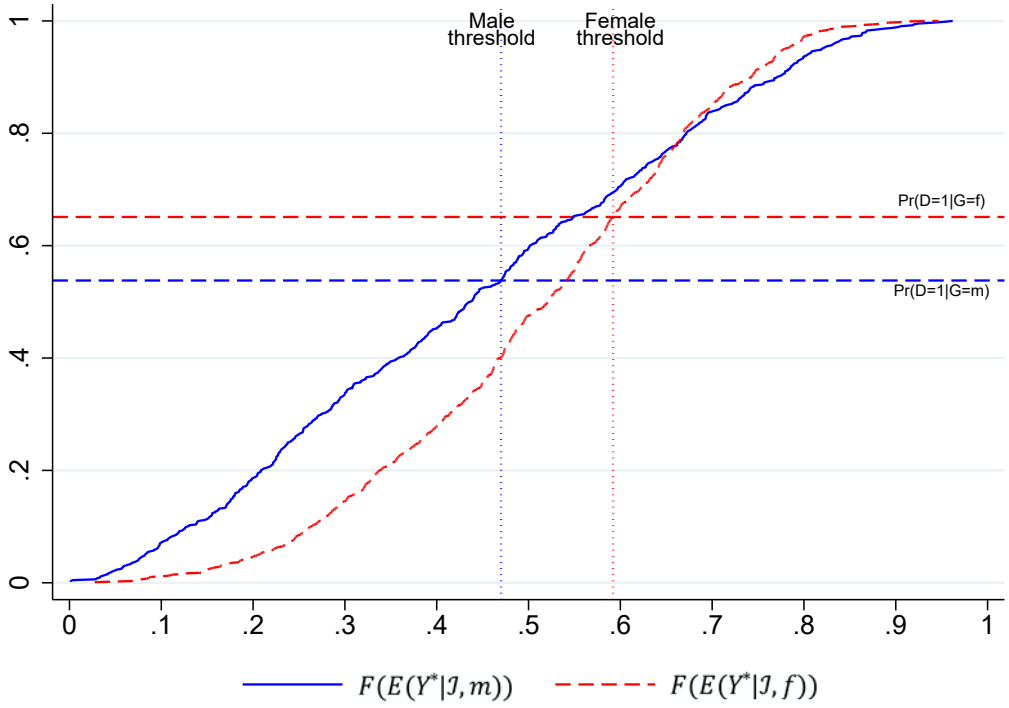

Figure 3: Identifying differences in τ_g^*.

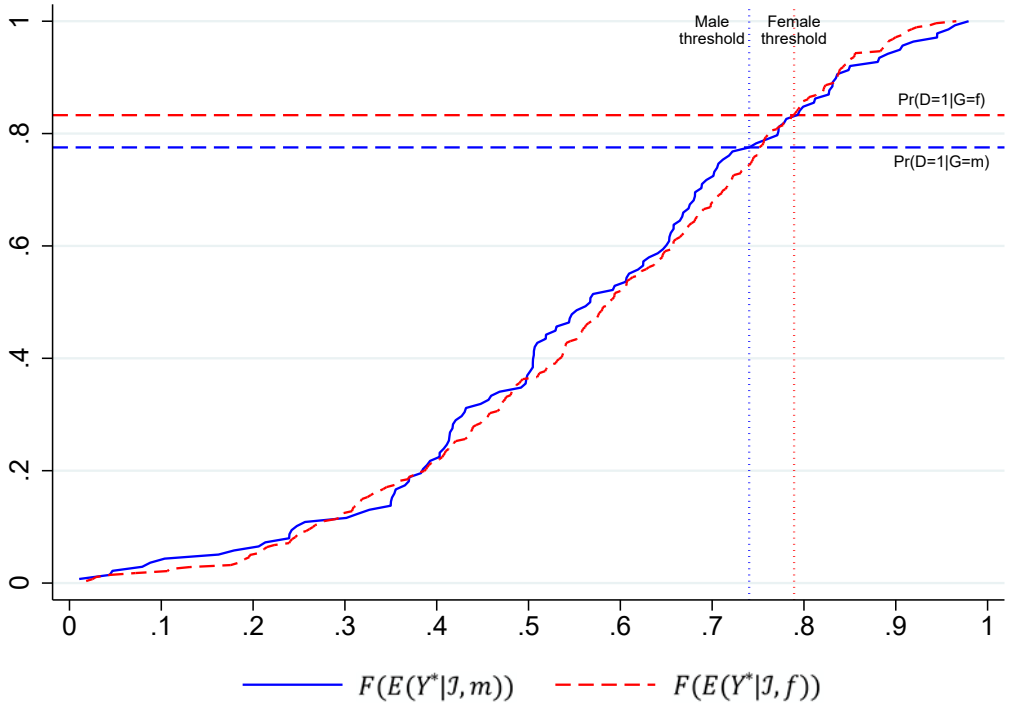

Figure 4: Identifying differences in τ_g^* at reconsideration.

Online Appendix

Table OA.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.

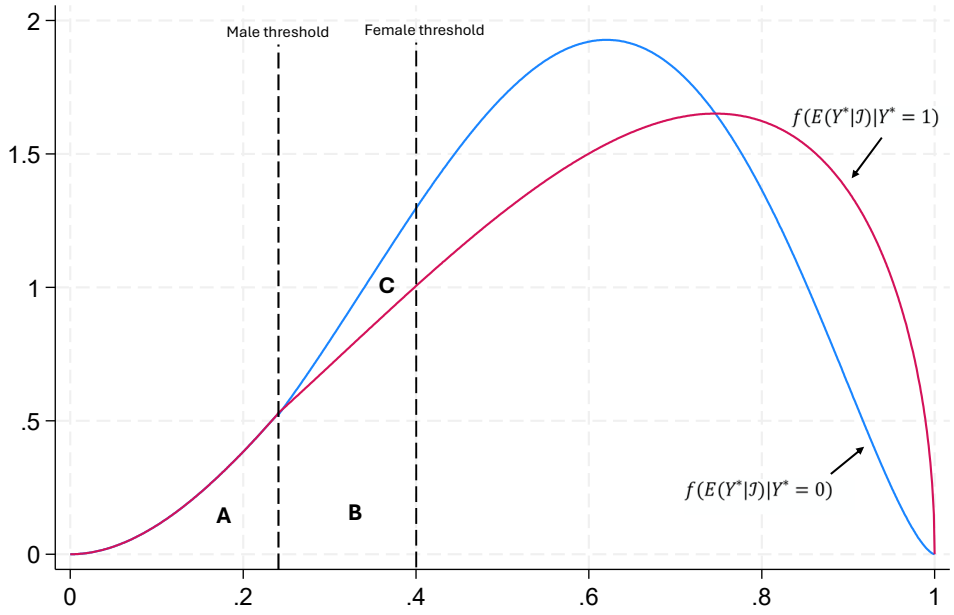

Figure OA.1: Failure of the average outcome test.

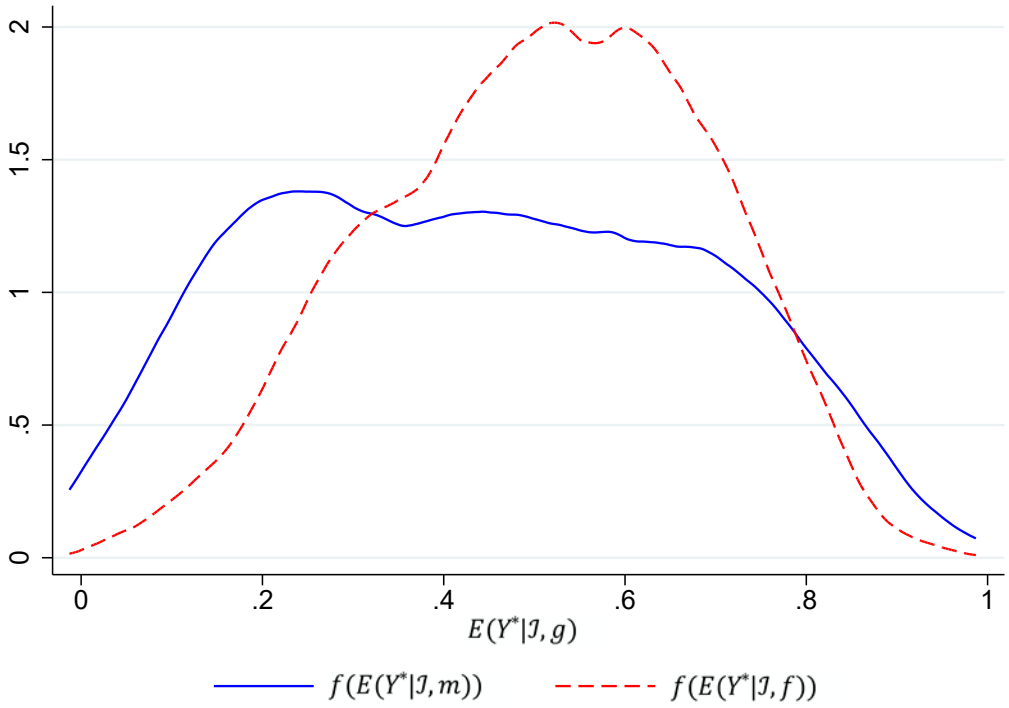

Figure OA.2: The empirical distribution of $E(Y^*|\mathcal{I}, G)$.