Evidence of Discrimination*

Nicola Persico
Northwestern University[†]

June 13, 2012

Abstract

This paper develops a decision-theoretic model of evidence production in the context of a discrimination trial. Producing evidence is assumed to be costly, and the cost can vary depending on what type of defendant behavior (and plaintiff characteristics) the evidence bears upon. The goal of the trial is to uncover a possible behavioral bias in the defendant (intent to discriminate). We then ask how a "social planner" would structure the production of evidence in a trial in order to best achieve this objective, taking into account the cost of evidence production. We show that it is sometimes efficient to sequence the production of different kinds of evidence (burden-shifting), or even to allow a decision based on limited evidence (e.g.., disparate impact alone, as a proxy for intent to discriminate). A key variable is the availability of evidence concerning the "productivity" of the plaintiff.

In this paper we lay out a simple decision-theoretic model of evidence production in a discrimination case. For this purpose, we take the perspective that the goal of discrimination law is to root out that behavior which is driven by a psychological bias against a protected class. This bias is conceptualized as a parameter in the defendant's objective function; in our analysis, this parameter will embody the legal concept of "intent to discriminate.[1]"

[*] This paper was written while in the employment of New York University. I am grateful to Thomas Miles for suggesting significant improvements in the paper, and to two anonymous referees. This research was partially supported by NSF research grant SES-0922381.

[†] Kellogg School of Management, 2001 Sheridan Road, Evanston, IL 60208. Email: nicola@nicolapersico.com

[1] In our analysis, intent to discriminate need not be a conscious choice on the defendant's part. It is sufficient that the defendant acts "as if" he were consciously taking into account the bias parameter.

We study the choices of a benevolent social planner who sifts through costly evidence in order to make inference about the unobserved bias parameter. The social planner weighs the probativeness of different pieces of evidence against the cost of producing them. This trade-off will sometimes lead the planner to optimally "economize" on evidence production. We will explore the hypothesis that the law is designed to mimick the choices of this social planner, and we will study the circumstances (parameter values in the model) in which it is optimal to economize on evidence.

Our analysis focuses on the probative value of different pieces of information. We focus on two different pieces of information: evidence about the defendant's behavior across different protected categories, and evidence about the heterogeneity in "productivity" across different categories of potential plaintiffs. When the social planner "economizes" on evidence, in our setting this means that he chooses to make a decision without information about productivity.

We connect the lessons from our model to actual evidentiary standards. We analogize choice based solely on the defendant's behavior, in the absence of information about productivities, to a standard of evidence based on **disparate impact**. Conversely, when the social planner optimally chooses to also acquire evidence about productivity, we interpret our model as calling for evidentiary standard based on **disparate treatment** (or intentional discrimination). Formally, of course, the disparate impact and disparate treatment are different causes of action, not different evidentiary standards. It seems reasonable, however, to operationalize them in a model as being different evidentiary standards.[2] Using this analogy, we explore the predictive power of our theoretical framework in explaining the variation in the behavior that is deemed illegal in various antidiscrimination statutes and laws.

Next we introduce the setup and discuss its virtues and limitations.

1 The setup

We have one defendant and N treated. The defendant is the agent suspected of discrimination. This agent applies a treatment to the treated. The treated are a pool of individuals who are affected by the defendant in some way; for example, they might be the set of employees to whom a defendant (employer) pays a wage, or motorists who are subject to search by the police. Each defendant is identified by an index i and a race $R \in \{A, W\}$. Here race also stands in for any protected category, such as age, gender, etc. All the treated receive $x_{iR} \geq 0$ from the defendant (in the labor discrimination context, a wage, in the police case, a probability of searching the motorist).

[2] Moreover, there is confusion in the law as to whether disparate impact protects employees from a conceptually distinct discriminatory injury than disparate treatment. See, for example, Ricci v. DeStefano, 129 S.Ct. 2658, 2682 (2009) (Scalia, J., concurring) (noting that the Court has not resolved whether disparate impact is merely "an evidentiary tool used to identify genuine, intentional discrimination—to 'smoke out,' as it were, disparate treatment"); George Rutherglen, Employment Discrimination Law 72-73 (2001).

There are N_R individuals of race R among the treated. In what follows, we implicitly assume that either we are looking at the totality of the treated population, or that we observe a representative sample of size N of the treated population.[3] The defendant may also have selected the treated population in a prior phase (for example, a case of wage discrimination takes as given the employee population, which of course results from the employer hiring decisions). In this case the analysis is limited to bias in the current phase, and not with bias in the prior "selection" phase.

In addition to these players there is a social planner who represents the judicial system or the interests of society. The social planner must convict or acquit the defendant based on the available evidence and on a set of rules to interpret the evidence. This set of rules is specified ex ante, before seeing the specific evidence for a given defendant. This set of rules is the ultimate object of our analysis.

1.1 Defendant's behavior

The defendant chooses the vectors $(x_{iR})_{i=1}^{N_R}$ to maximize

$$\sum_{i,R}\left[x_{iR}\cdot(\gamma_{iR}-\beta_{iR})-\frac{(x_{iR})^2}{2}\right]. \tag{1}$$

The term γ_{iR} represents an objective "productivity" pertinent to agent (i,R). What this productivity corresponds to in reality will be made clear in the examples that follow. The term β_{iR} represents a subjective bias term that affects the defendant's objective function; a large β_{iR} is perceived behaviorally by the defendant as lowering the productivity of agent (i,R). The defendant knows γ_{iR},β_{iR} when choosing x_{iR}.[4] The negative squared term in this payoff function represents the cost to the defendant of action x_{iR} (for example, the opportunity cost of moneys paid in wages).

Although very stylized, this model captures certain crucial elements in many discrimination settings. A few examples follow.

Example 1 ***Wage discrimination:*** *Let x_{iR} denote the wage rate and $2(\gamma_{iR}-\beta_{iR})$ the employee's perceived productivity (objective productivity minus bias factor). Suppose the employee chooses effort e to maximize his salary net of the cost of effort, $\max_e x_{iR}e-\frac{e^2}{2}$ so that the optimal effort choice is $e^*=x_{iR}$. Then the perceived revenue generated by the employee is given by $e^*2(\gamma_{iR}-\beta_{iR})=x_{iR}2(\gamma_{iR}-\beta_{iR})$, and the firm's wage bill is $e^*x_{iR}=(x_{iR})^2$. The*

[3] The analysis is not valid if we observe a selected sample of the treated population, say, those who receive the worst treatment.

[4] Of note, the defendant here need not make any inference regarding the productivity of each treated, since the γ_{iR}'s are assumed to be known. In this model, therefore, the complicated issue of "statistical discrimination" does not arise.

firm maximizes perceived revenue minus labor costs. Divide through by 2 to get the objective function (1).

Example 2 ***Hiring discrimination:*** *x_{iR} is the probability of hiring the applicant, γ_{iR} the applicant's productivity. The square term in the objective function captures the opportunity cost of the resource x.*

Example 3 ***Lending discrimination:*** *x_{iR} is the amount of money lent to the the applicant, γ_{iR} the applicant's probability of repaying.*

Example 4 ***Discrimination in police searches:*** *x_{iR} is the probability of searching the motorist, γ_{iR} the motorist's likelihood of carrying illicit drugs.*

Example 5 ***Selective prosecution:*** *x_{iR} is the probability that the prosecutor prosecutes the case, γ_{iR} the likelihood that the case results in a conviction (a victory for the prosecution).*

For each i, R the defendant solves

$$\max_{x_{iR}} x_{iR} \cdot (\gamma_{iR} - \beta_{iR}) - \frac{(x_{iR})^2}{2}.$$

This is a concave function of x_{iR}, and the first-order conditions are

$$x^*_{iR} = \gamma_{iR} - \beta_{iR}. \tag{2}$$

These first-order conditions uniquely identify the maximum, provided the maximum is interior. To ensure interiority of x^*_{iR} we will assume henceforth that $\gamma_{iR} > \beta_{iR}$. This assumption guarantees that all treated receive a strictly positive x^*_{iR}.[5]

The objective function (1) does not reflect considerations pertaining to the possibility of being scrutinized, and possibly punished, for being found guilty of bias. In so doing we ignore the endogeneity of the discriminator's behavior to the social planner's policies (i.e., the rules used by the courts). This is an important "partial equilibrium" assumption which allows us to deal with the heart of our matter (rules for evidence production) without having to worry about the feedback of these rules on the discriminator's behavior. From a formal perspective, our analysis approximates an environment in which the probability of beign sued for discrimination is perceived as small by the discriminator. We believe that not much would change qualitatively in our analysis were we to incorporate this feedback effect.[6]

[5] Without this assumption we have a censoring problem, which can be dealt with using specific statistical techniques.

[6] That being said, we feel that endogeneizing the discriminator's response to the probability of being tried would be an important direction of future work.

Equation (2) shows that the defendant's behavior towards the treated is the sum of two components: the treated's ability/productivity, and the defendant's bias towards the treated. Now we specify the process which, for each treated, generates a productivity and a bias.

Assumption 1 *Each individual productivity γ_{iR} is an independent realization from the random variable Γ_R. Each individual bias factor β_{iR} is an independent realization from random variable B_R. We assume that $\Gamma_R - B_R > 0$.*

This assumption says that the productivity of each individual in our sample varies, as does the bias factor which clouds the defendant's view of that particular individual. The assumption implies that, within race, bias and ability are not correlated. That is, the defendant is not systematically more biased against more (or less) able individuals. The last bit of the assumption ensures that $\gamma_{iR} > \beta_{iR}$ for all i, R. This assumption is needed to avoid the case $x^*_{iR} = 0$.

Note that we are not making any functional form assumption on B_R, Γ_R. Note also that we allow $\Gamma_A \neq \Gamma_W$, that is, the distributions of productivity is allowed to differ across protected categories. This is an important feature of our analysis, since we want to allow for the possibility that what appears like discrimination is in fact merely a reflection of an unequal distribution of abilities.

1.2 The social planner's objective

The social planner represents the judicial system, or the interests of society as a whole, as pertains to detecting whether the defendant is biased. To make this objective formal we need to specify it mathematically. Let's start with some notational conventions. For a generic random variable X we denote its expected value by μ_X and its variance by σ^2_X. We denote the difference in bias across races as $B = B_A - B_W$, and the difference in productivities across races as $\Gamma = \Gamma_A - \Gamma_W$. Now we specify what it means to be biased. We say that a defendant is biased if the probability distributions from which the treated-specific biases are drawn have different means.

Definition 1 *A defendant is biased if $\mu_B \neq 0$.*

This is of course a very specific formalization of what it means for a defendant to be biased against protected groups.[7] This definition says that a defendant is biased if he behaves toward the treated as if, *in expectation*, he discounts one group's productivity more. This formalization seems like a reasonable starting point for our analysis.

[7] In particular, this formalization abstract from any other differences between the two distributions save for their first moments.

A notable feature of our formalization is that a defendant's biases against each individual member of race R are drawn independently from a probability distribution. This formulation makes it possible for a defendant to be biased against one member of race R and not biased against another member of the same race. Our model, of course, allows as a special case the one in which the bias is effectively the same across all members of race R. This special case obtains when the variance of the random variable B_R is very small (or even zero), which implies that the draws β_R will all be very close to each other.

We now place additional structure on the means of the two distribution.

Assumption 2 *With prior probability π we have $\mu_B = K_B$, and with complementary probabilty $\mu_B = 0$.*

π is the prior probability that the defendant is biased, that is, the subjective probability in the social planner's mind, before any evidence has been introduced, that for this defendant $\mu(B_A) \neq \mu(B_W)$ .This assumption places structure on the prior (subjective) probability that the social planner gives to the average relative bias. This assumption says that either the defendent is perfectly unbiased ($\mu_B = 0$) or else he has a bias of a "size" equal to exactly K_B. Clearly one could have been more general and allow for a richer set of possibilities. However, this simple formulation is sufficient to bring out the forces that govern our problem. One may wish to impose normatively that $\pi = 1/2$, in order to capture an "unprejudiced" social planner.

The social planner needs to take an action, acquit or convict, not knowing for sure whether the defendant is in fact biased. The social planner's payoff from taking the action is given by the following expected loss function.

$$L = (1-\pi)\, l_{CI} \Pr(CI) + \pi\ l_{AG}\ \Pr(AG). \tag{3}$$

We take this expression to be the social objective in a discrimination trial. Here l_{CI} and l_{AG} are exogenously given positive parameters, which capture the loss from a mistake: convicting the unbiased (innocent), and acquitting the biased (guilty), respectively. The loss associated with making the correct decision is assumed to be zero.[8] One may wish to assume that the ratio l_{CI}/l_{AG} is very large, to capture the judicial system's concern for convicting of discrimination defendants who are, in fact, unbiased. Of course our formalization allows for this parameter configuration, and for many others. $\Pr(CI)$ and $\Pr(AG)$ represent the probability of making these mistakes, and they depend on the the social planner's decision process. In our intepretation, these probabilities capture the standard of proof required to convict, what type of evidence is considered, the specific form of the doctrinal test applied, etc.). We will focus on these probabilities shortly.

[8] In this simple formalization the loss does not depend on the egregiousness of the discrimination. This is done for simplicity.

1.3 The available evidence

We introduce some assumptions about what the social planner knows and what he can learn at a cost.

Assumption 3 *The social planner knows μ_Γ but not μ_B.*

The assumption that μ_Γ is known means that there is no uncertainty in the mind of the social planner about the differences across races in *average* productivity. But even though the expected value of the distribution Γ is known, individual productivities in our sample are still randomly drawn, and thus unknown. They represent a powerful confounder in the social planner's inference problem if he is trying to recover bias from observing x^* alone (this is seen formally in equation 2).

Assumption 4 *The social planner knows the variances of all random variables.*

This assumption is merely technical and it simplifies the exposition. If we did not know these variances, they would be proxied for by the empirical variances.

We now describe what information is available to the social planner. In our interpretation, this is the type of evidence that can be introduced in a discrimination trial. We do not yet address how this evidence is to be used to form the probabilities $\Pr(CI)$ and $\Pr(AG)$.

Assumption 5 *The social planner can learn $\frac{\sum_i x^*_{iR}}{N_R}$ (average treatment for race R) at a cost C_β, and $\frac{\sum_i \gamma_{iR}}{N_R}$ (average productivity in race R) at a cost C_γ.*

We interpret learning $\frac{\sum_i x^*_{iR}}{N_R}$ as producing evidence through discovery about the average treatment of treated in race R. We will interpret learning $\frac{\sum_i \gamma_{iR}}{N_R}$ as producing evidence through discovery about the average productivity of treated in race R. For example, in the context of wage discrimination x^*_{iR} represents the wage earned by worker (i, R) and γ_{iR} represents the productivity of that worker. In the context of racial profiling, $\frac{\sum_i x^*_{iR}}{N_R}$ represents the fraction of motorists of race R who are searched among the N_R stopped motorists, and $\frac{\sum_i \gamma_{iR}}{N_R}$ represents the fraction of stopped motorists who carry drugs. Note that we assume that the productivity of all the treated can be discovered, irrespective of their level of treatment x^*_{iR}. This assumption greatly simplifies our analysis.[9]

[9] An alternative assumption, which we do not entertain, would be that we do not observe $\frac{\sum_i \gamma_{iR}}{N_R}$ but rather $\frac{\sum_i x^*_{iR}\gamma_{iR}}{N_R}$. This statistic corresponds to the equilibrium "return" to the treated from treating race R, and is what is used in the so-called "outcome tests." (See Ayres 2002). Of course, one may not need to worry about outcome tests in our context so long as $x^*_{iR} > 0$, since in that case we can back out γ_{iR} from knowledge

The costs C_β and C_γ can be thought of as the social costs of producing evidence through discovery. In practice, these costs can vary significantly across different contexts. For example, in the context of police searches, evidence on at least the x^*_{iR} (probability of being searched) can be very easy to produce (low C_β) if there is a record of all stops and searches; moderately difficult (medium C_β), if there is a record of searches but there is no clear data on the entire population for which we are trying to compute this probability;[10] or very difficult (high C_β) if no records of searches are kept. Beyond the direct costs of producing evidence, there are indirect costs. For example, a firm may not want to disclose how much it pays its employees.

Assumption 5 implies that the social planner can exactly learn the x^*_{iR}'s and γ_{iR}'s. This is a stark simplification of reality. In many practical cases it is not easy to know exactly how the defendant is being treated (the x^*_{iR}'s). It is even harder to know the defendant's productivity withour error (the γ_{iR}'s). For example, outcome measures (probability of finding illicit drugs in a motor-vehicle search, or measures of productivity for an employee) are mere proxies for the γ_{iR}'s. Still, even if we only have proxies for the x^*_{iR}'s and γ_{iR}'s, the analysis outlined below can be carried out using the proxies, and appropriately accounting for this fact when computing the variance of certain statistics.

1.4 The social planner's problem

The social planner's chooses:

a) what type of evidence to evaluate (x^*_{iR}'s and γ_{iR}'s), if any, and

b) what standard of proof should be applied to the evidence,

in order to minimize the expected loss function (3) plus the cost of evaluating evidence.

2 Optimal standard of proof for given evidence type

In this section we ask what the optimal standard of proof should be, for given type of evidence available to the planner. The optimal standard of proof minimizes the loss function of equation (3). In practice, the optimal standard of proof will take the form of a cutoff applied to a statistic (a summary of the type of evidence available), such that the cutoff will identify a "conviction region" and an "acquittal region." This cutoff, together with the

of x^*_{iR} and $x^*_{iR} \cdot \gamma_{iR}$. However, in many cases of practical relevance $x^*_{iR} = 0$ is a frequent occurrence. For instance, in the case of racial profiling we may not observe the probability of treating x^*_{iR}, but rather only its realization $\widehat{x}^*_{iR}$, which is either 1 (searched) or zero (not searched). Therefore, when $\widehat{x}^*_{iR} = 0$ observing $\widehat{x}^*_{iR} \cdot \gamma_{iR}$ is uninformative about γ_{iR}. In words, the productivity (whether drugs are found) of a motorist who is not searched is usually unknown. In such cases, recovering the γ_{iR}'s from observed outcomes (productivity of treatment) is much more challenging. In this paper we will not address this subject.

[10] This is a problem because the population goes in the denominator in order to compute the probability of being searched.

statistic to which it is applied,will determine the functional form of the terms $P\left(CI\right)$ and $P\left(AG\right)$ and thus the expected loss.

2.1 Optimal standard of proof under a disparate treatment test

In this section we characterize the functional form of the terms $P\left(CI\right)$ and $P\left(AG\right)$, in the case in which $\frac{\sum_i x^*_{iR}}{N_R}$ (average treatment for race R) and $\frac{\sum_i \gamma_{iR}}{N_R}$ (average productivity in race R) are both observable. For given R, we choose to combine these quantities in the statistic

$$T_{1R} = \frac{\sum_i \left(\gamma_{iR} - x^*_{iR}\right)}{N_R}$$

This statistic is not merely a disparate impact test because it incorporates more than just evidence about treatment (the x^*_{iR}'s). Indeed, in this statistic evidence about treatment is "benchmarked" against evidence about productivity (the γ_{iR}'s). Put differently, this statistic is large when members of race R are treated worse than their productivity warrants. In fact, this statistic corresponds to a disparate treatment test in the sense that it is directly informative about a defendant's bias μ_{B_R}. Indeed, equation (2) allows us to rewrite

$$T_{1R} = \frac{\sum_i \beta_{iR}}{N_R}.$$

In words, under the model of defendant behavior which gives rise to equation (2), T_{1R} is the sample mean of the β_{iR}'s which, recall, are realizations from the random variable B_R. Therefore, this statistic is an estimator for μ_{B_R}. In fact, as the sample gets large T_{1R} converges in distribution to a normal with mean μ_{B_R} and variance $\frac{\sigma^2_{B_R}}{N_R}$. This converges is written compactly as

$$T_{1R} \rightsquigarrow N\left(\mu_{B_R}, \frac{\sigma^2_{B_R}}{N_R}\right). \tag{4}$$

Therefore not only is T_{1R} is informative about μ_{B_R}, but moreover we recover the (asymptotically Normal) distribution of T_{1R} despite making no assumption on the distribution B_R.

This information can be put to use to obtain precise functional forms for $\Pr\left(CI\right)$ and $\Pr\left(AG\right)$ and, therefore, for the loss function. To this end, suppose that the planner uses the following threshold rule: convict iff $T_{1A} - T_{1W} > t_1$, where t_1 is a threshold which later will be chosen optimally. Intuitively, we convict if the disparity between average abilities and average treatment is greater in the A population than in the W population. Then we can give a precise asymptotic expression for the loss function.

Proposition 1 *Suppose that the planner uses a disparate treatment test, and thus convicts if and only if $T_{1A} - T_{1W} > t_1$. Then asymptotically the expected loss as a function of t_1 is* *given by*

$$L_1(y) = (1-\pi)\, l_{CI}\, [1-\Phi(y)] + \pi\ l_{AG}\ \Phi\left(y - \frac{K_B}{\sqrt{\frac{\sigma^2_{B_A}}{N_A} + \frac{\sigma^2_{B_W}}{N_W}}}\right), \tag{5}$$

where $y = \frac{t_1}{\sqrt{\frac{\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{B_W}}{N_W}}}$ *and* Φ *denotes the c.d.f. of a Normal distribution with mean zero and variance equal to 1 (standard Normal). The function* $L_1(y)$ *is concave in* y *and has an interior minimum* y^**. The optimal standard of proof is* $t_1^* = y^*\sqrt{\frac{\sigma^2_{B_A}}{N_A} + \frac{\sigma^2_{B_W}}{N_W}}$*. The optimum* y^**, and therefore the optimal standard of proof* t_1^**, is increasing in* $\frac{(1-\pi)l_{CI}}{\pi\ l_{AG}}$*.*

Proof. See the appendix. ■

This proposition gives an expression for the loss associated with any standard of proof that can be applied to a disparate treatment test. A large value of t_1 corresponds to a high (demanding) standard of proof required to convict the defendant. The optimal standard of proof trades off two types of mistake: As y grows, the first addend in the loss function becomes smaller. This is because we are convicting less frequently, and so the risk of convicting the innocent becomes smaller. The second addend instead grows with y, because as we convict less frequently it is more likely that the guilty is acquitted. The optimal threshold, or standard of proof, is the one that optimally trades off these two types of mistakes.

2.2 Optimal standard of proof under a disparate impact test

In this section we derive a functional form for the loss function under the assumption that the planner only sees the x^*_{iR} but not the γ_{iR}'s. In our interpretation this corresponds to a disparate impact test, because the planner ignores evidence about differential productivities.

In this case the natural statistic to use is

$$T_{0R} = \frac{\sum_i x^*_{iR}}{N_R}.$$

Using (2) we have

$$T_{0R} = \frac{\sum_i \gamma_{iR}}{N_R} - \frac{\sum_i \beta_{iR}}{N_R} \rightsquigarrow N\left(\mu_{\Gamma_R} - \mu_{B_R}, \frac{\sigma^2_{\Gamma_R} + \sigma^2_{B_R}}{N_R}\right). \tag{6}$$

This statistic corresponds to a disparate impact test in the sense that it measures how differently two groups are treated on average, but it does not account for potential differences in productivity between these groups.

The limiting distribution of T_{0R} should be compared with the limiting distribution of T_{1R} (see 4). Whereas the latter is purely a function of B_R, the distribution we want to make inference about, the distribution in (6) is "contaminated" by the presence of terms referring to Γ_R. Fortunately the term μ_{Γ_R} is known by assumption, and so it can be factored out. However, the variance of T_{0R} is seen to be greater than the variance of T_{1R}. The difference is due to the presence of the term $\sigma^2_{\Gamma_R}$, which captures the "added noise" due to the fact that T_{0R} varies together with $\sum_i \gamma_{iR}$. Intuitively, looking at treatment only (the x^*_{iR}'s) conflates two elements: one that we care about, the possible bias; and a confounding factor that we do not care about, the unobserved variation in productivities, which we cannot eliminate.[11]

Following the same steps as in the previous section, we form the statistic $T_{0W} - T_{0A}$. We will consider a procedure that convicts when this statistic exceeds a threshold t_0. Intuitively, this is an event in which group W is on average overcompensated relative to group A.

Proposition 2 *Suppose that the planner convicts if and only if $T_{0W} - T_{0A} > t_0$. Then asymptotically the loss function is given by*

$$L_0(y) = (1-\pi)\, l_{CI} \left[1-\Phi(y)\right] + \pi \ l_{AG} \ \Phi\left(y - \frac{K_B}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A} + \frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}}\right), \qquad (7)$$

where $y = \frac{t_0+\mu_\Gamma}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A} + \frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}}$ and Φ denotes the c.d.f. of a Normal distribution with mean zero and variance equal to 1 (standard Normal). The function $L_1(y)$ is concave in y and has an interior minimum y^*. The optimal standard of proof is $t^*_1 = y^* \sqrt{\frac{\sigma^2_{B_A}}{N_A} + \frac{\sigma^2_{B_W}}{N_W}}$. The optimum y^*, and therefore the optimal standard of proof t^*_1, is increasing in $\frac{(1-\pi) l_{CI}}{\pi \ l_{AG}}$.

Proof. See the appendix. ■

3 Best tests

In this section we consider the problem of a planner who is considering whether to acquire evidence about treatment or productivity or both. We ask what kind evidence the social planner would want to base his decision on. We interpret different kinds of evidence as corresponding to different doctrinal tests, or analyses, for establishing discrimination. We will consider two different kinds of tests: conditional and unconditional tests.

[11] Or, more precisely, which we can only eliminate "on average."

3.1 Unconditional tests

In this section we consider unconditional tests, which means that we do not allow tests such as: acquire information about treatment and, if after looking at that evidence it seems like there is cause for concern, then also acquire information about productivities. Examples of unconditional tests are those based on disparate impact, and those based on disparate treatment (intentional discrimination).

The first observation is that information on productivity alone is useless to the social planner in making inference about μ_B. So we only need to consider three evidence configurations: $\{\varnothing\}, \{x^*_{iR}\}$, and $\{\gamma_{iR}, x^*_{iR}\}$. The first configuration denotes the case in which the decision is based on no evidence, the second to the case in which the decision is based only on evidence on treatment (disparate impact standard), the third to the case in which the decision is based on a disparate treatment (intent to discriminate) standard. There first configuration (no evidence) is not interesting, so from now on we focus on the case in which that evidence configuration is never optimal. The exact conditions on primitives required to rule out this case are presented in Appendix B.

The value of evidence about productivities is given by its effect on the expected loss that can be achieved using that evidence. difference between the values of problems (8) and (7). Call this value $V(\gamma)$, so

$$V(\gamma) = L_0(y^*_0) - L_1(y^*_1)$$

Proposition 3 *Suppose the planner must choose between two unconditional tests: disparate impact, and disparate treatment. Then*

a. The disparate treatment test is optimal if and only if $C_\gamma \leq V(\gamma)$.

*b. Which of the two tests is optimal does not depend on the heterogeneity in average productivity **across** protected categories (μ_Γ).*

*c. The higher the heterogeneity in productivity **within** protected categories ($\sigma^2_{\Gamma_A}$ and $\sigma^2_{\Gamma_W}$), the more likely it is that the planner chooses the disparate treatment test.*

d. For any parameter configuration there is always a C_γ sufficiently large such that the disparate treatment test is not optimal, and a C_γ sufficiently small that it is optimal.

Proof. See the appendix. ■

3.2 Conditional test (McDonnel-Douglas)

A conditional test is a rule that calls for acquiring evidence on productivities γ_{iR} only after observing certain realizations of treatment $T_{0W} - T_{0A}$. An example is the “McDonnel-Douglas" burden shifting analysis, whereby first the treated is required to prove disparate impact, and only if this step is successful is the treated required (and allowed) to present evidence about productivities that might justify the disparity in treatment. We will further explore this analogy below.

Definition 2 *A sequential test is a triple of thresholds $\underline{t_0}, \overline{t_0}, t_1^S$ such that: if $T_{0W} - T_{0A} < \underline{t_0}$ the defendant is acquitted. If $T_{0W} - T_{0A} > \overline{t_0}$ the defendant is convicted. If $T_{0W} - T_{0A} \in \left[\underline{t_0}, \overline{t_0}\right]$ then evidence on productivities γ_{iR} is acquired, and then the defendant is convicted if and only if $T_{1A} - T_{1W} > t_1^S$.*

When will the sequential evidence verification rule prescribe that evidence on productivities γ_{iR} be acquired? Intuitively, when the value of additional information is higher. This happens for realizations of $T_{0W} - T_{0A}$ in their “middle range.” The exact location of this middle range depends on the costs of type I and type II errors, among other parameters. When these two costs are of similar size, then the middle range lies around $-\mu_\Gamma$. When the realization of $T_{0W} - T_{0A}$ is extreme, either much above or much below $-\mu_\Gamma$, then there is a strong inference either that $\mu_B = K_B$ or that $\mu_B = 0$. In this case there is not much value in learning γ_{iR}. In contrast, information about γ_{iR} is most valuable when the realization of $T_{0W} - T_{0A}$ is middling, because in this case we cannot be sure about the value of μ_B.

Notice that unconditional tests are a limit case of the sequential test. Indeed, the disparate impact test obtains when $\underline{t_0} = \overline{t_0}$, and the disparate treatment test obtains when $\underline{t_0} = -\infty, \overline{t_0} = +\infty$. Therefore, the sequential test must always be at least weakly preferred by the social planner. In fact, the social planner's preference will frequently be strict. That is, the social planner will usually strictly prefer to proceed sequentially rather than unconditionally. The reason is twofold. First, suppose that among unconditional tests the planner prefers a disparate impact test to a disparate treatment test. This must be because the cost C_γ is too high *ex ante*, that is, averaging over all possible realizations of $T_{0W} - T_{0A}$, the added information contained in the productivity is not worth the cost of acquiring that information. But even so, it can still be true that *conditional on some realizations of* $T_{0W} - T_{0A}$, namely those around $-\mu_\Gamma$, the added information is worth the cost. In this case preference will be strict. Conversely, suppose that among unconditional tests the planner prefers a disparate treatment test. This must be because the cost C_γ is low enough that, *for most realizations* of $T_{0W} - T_{0A}$, the added information contained in the productivity is well worth the cost of acquiring that information. But even so, no matter how small C_γ, for N_R large enough there will be realizations of $T_{0W} - T_{0A}$ so extreme (very large or very small) that the relative likelihood of innocence and guilt conditional on those realizations is arbitrarily close to zero or 1. In these events it is very unlikely that additional information about productivities will change that ratio enough to change the decision. Therefore it is optimal for the planner to forgo this information. The following proposition summarizes this discussion.

Proposition 4 *Fix all parameters except C_γ. There is a value $\overline{C_\gamma} > V_\gamma$ such that: for $C_\gamma < \overline{C_\gamma}$ the social planner strictly prefers the sequential test to any unconditional test; and for $C_\gamma > \overline{C_\gamma}$ the social planner strictly prefers the disparate impact test to any other test, conditional or sequential.*

This proposition says that in this setting there is no role for a disparate treatment test in this setup.

4 The incentives of the parties to the trial

The analysis developed in the previous sections can account for two tests: a disparate impact test, which is optimal when the cost of acquiring information about productivities is high, and a conditional test which is optimal in all other circumstances. In this section we turn to a different setup, one in which there is no social planner who can choose when to initiate the (various phases of) the tests. Instead, there are two parties, plaintiff and defendant, who strategically choose whether to go forward in a discrimination trial, given the doctrinal tests that the law imposes.

A reason why we might want to explore this more strategic setup is that the analysis developed above provides no role for a disparate treatment test: this test is dominated by the sequential (burden-shifting) test. But the disparate treatment test is very common in reality, and so we want to know whether the strategic formulation can explain its prevalence.

The new setup is as follows. First, the parties in a lawsuit have incentives that are not aligned with the social planner's. The plaintiff and the defendant do not care about the probability of making the wrong decision, but rather about the probability of winning the lawsuit. Second, the plaintiff typically does not fully internalize the cost of producing evidence (the process of discovery, for example, can be very expensive for the defendant). The social planner (the judge, or the judicial system) has two sets of instruments to govern the parties' actions: one is the type of evidence that he requires before making a decision. The second instrument is the standard of proof applied to the evidence.

First, we demonstrate that if plaintiff and defendant can be persuaded to go through the various steps of the lawsuit, then a sequential test can be implemented by shifting the burden of proof, i.e., by a McDonnel Douglas type of analysis. To see this, consider the the first leg of MDD and set the standard of proof in the first leg at at $\underline{t_0}$. If $T_{0W} - T_{0A} < \underline{t_0}$, then this threshold is not met then the defendant is acquitted. If the threshold is met, then the defendant is allowed (but not required) to present as a defense evidence about productivities. The optimal standard of proof for this second leg of the test will be set at t_1^S. Thus we can replicate any sequential test, *provided* that the defendant opts to go through to the second leg of the test if and only if $T_{0W} - T_{0A} < \overline{t_0}$. However, there is no guarantee that the defendant will follow this strategy. If the defendant is insufficiently motivated to “fight” the second leg of the MDD test, then it might not be possible to implement a sequential test in an environment with strategic parties simply by shifting the burden of proof. It might then be preferable to “force” the production of evidence on productivities, which can be done by settling for an intentional discrimination (disparate treatment) test. In this context the advantage of a disparate treatment test over a burden-shifting test is that the plaintiff has the power to compel the defendant to produce evidence on productivities (through discovery).

5 Summary of predictions

Our analysis suggests that a disparate impact standard should prevail when the cost of acquiring information about productivities is large. When that cost is moderate or low for at least one party then we should expect either a MDD standard or a disparate treatment standard. We should expect the latter if the defendant has insufficient motivation (relative to the social optimum) to fight the charges, whereas the former should prevail when the defendant has an advantage in the production of evidence on productivities.

A different consideration, and a reason why the disparate treatment test might be dominated by the sequential test is that in the disparate treatment test the burden of proof of intentional discrimination rests with the plaintiff. In the sequential test (the MDD test), in contrast, the burden of proof rests with the defendant. Thus in the disparate treatment case the plaintiff may need to expend considerable costs interpreting the evidence about productivity (perhaps provided by the plaintiff through the discovery process). This is a social cost which may be avoided if the burden of proof falls on the defendant. Usually, the defendant will be able to interpret evidence about productivity more easily than the plaintiff, who might not be cognizant about the nuances of the production process. In our model, this wedge can be introduced by making the cost C_γ actor-specific: we can then allow the defendant to have a lower cost than the plaintiff. Under this view, the advantage of the MDD test over the intentional discrimination test is that it places the burden of producing proof on the party that can produce the proof for cheaper.

6 Case law

6.1 Wage discrimination

Wage discrimination cases are regulated by a specialized body of case law which has crystallized in the so-called McDonnel-Douglas analysis. According to this analysis, the plaintiff must first establish disparate impact, and then the burden shifts on the defendant to disprove intent to discriminate. It is reasonable to suppose in this case that the cost of acquiring evidence about productivity is lower for the defendant than for the plaintiff, given the employer's interest in monitoring productivity in the course of day-to-day operations. Moreover, in this case the defendant has considerable proprietary knowledge about the "production function," and so an outsider may find it difficult to determine exactly how to measure individual productivity. In other words, the cost of producing evidence about productivity is lower for the defendant. Therefore our theory suggests that a sequential test with burden-shifting is appropriate for this case. This, of course, is exactly the MDD test.

This observation must be qualified by the fact that the standard of proof in the second leg of the MDD test is "business necessity," which is a very stringent standard. In practice, the application of this standard tends to de-emphasize the role of individual productivities as a valid defense for disparate impact. As a result, in many practical cases the MDD test becomes somewhat similar to a disparate impact test.

6.2 Selective enforcement

Selective enforcement and selective prosecution claims are regulated by the 14th amendment, and therefore they are formally subject to a disparate treatment test. However, according to our analysis selective enforcement should be divided into two categories. The first category is that in which few measures of productivity are generated during the enforcement process. One could think for example of enforcing fairness in the application of complex administrative regulations, where it is sometimes difficult to measure whether the regulation has been applied in a color-blind way. In this case C_γ is high, and so disparate impact may be the only possible test. The case of Yick Wo v. Hopkins, provides an interesting practical example. The U.S. Supreme Court argued that a San Francisco ordinance prohibiting the operation of laundries in wooden buildings had been applied selectively. On the basis of this ordinance a Chinese owner was forbidden from operating a laundry business in a wooden building, whereas eighty non-Chinese owners were allowed to operate laundries in wooden buildings. It is entirely possible that the eighty non-Chinese laundries were operated in safe (though wooden) buildings, while the Chinese laundry was operated out of a patently unsafe building. Yet, in the absence of clear measures of "safety" of an operation, intent to discriminate was inferred from egregious disparate impact.

A different case is that in which the enforcement activity naturally generates some measure of productivity. The prototypical example are claims of racial discrimination in traffic stops. In the case of police searches, measures of productivity are potentially available even for the individual officer, through an outcome analysis. That is, it is possible to get a sense of the productivity of searching motorists of different races by looking at the frequency with which drugs are found on the motorist (success rate of searches). The exact analysis required to go from this statistic to a measure of intent to discriminate is subtle, and is developed in Knowles et al. 2001. However, the main point is straightforward: in this case we have evidence, albeit indirect, about productivities. Therefore, in this case the cost C_γ of acquiring evidence about productivity is not excessive, and in fact it is low provided that records are kept. Moreover, the defendant (police department) has no particular advantage in interpreting success rates, compared to the plaintiff. Therefore the cost of producing evidence about productivity is not much lower for the defendant. Therefore our theory suggests that a disparate treatment standard is appropriate for this case.

Of course, this hinges on there being statistics available about the productivity of searches. If the police departments do not keep such statistics then the cost C_γ is very high, and in this case a disparate impact standard would be preferable.

6.3 Jury selection

The case of possible bias in peremptory challenges (voir dire) refers typically to the scenario where a prosecutor might preferentially strike minority jurors from a jury. Bias in voir dire is governed by Batson v. Kentucky, 476 U.S. 79 (1986). That case establishes a sequential test analogous to the MDD test. According to that test, the defendant must first establish a case of disparate impact by showing that minority jurors have been struck disproportionately. If this first prong of the test is met, then the prosecutor can give his reasons for excluding such jurors. If the reasons dispel the assumption of intentional discrimination, then the prosecutor is allowed to strike the jurors. Of course, the mechanism by which a prosecutor decides to strike a juror is murky, and only the prosecutor himself can clarify the reasons. In other words, the prosecutor has considerable proprietary knowledge about the "production function," and so an outsider may find it difficult to determine exactly how to measure individual productivity. Since the cost of producing evidence about productivity is lower for the prosecutor, our theory suggests that a sequential test with burden-shifting is appropriate for this case. This is, of course, exactly the Batson test.

6.4 Lending discrimination

Nominally, lending discrimination is subject to a sequential, MDD-type test (Walter 1995). However, the federal agencies charged with enforcing fair lending have set a "business necessity" standard for the second leg of the test (Walter 1995, p. 67). This is a more stringent standard than the "legitimate business reason" test. In practice, this demanding standard has the effect of driving the enforcement of lending discrimination towards a disparate impact standard. However, the lending process automatically produces evidence about productivities (probability of repaying the loan), namely, the default rate. This evidence is available at very low cost (low C_γ). Moreover, a comparative analysis of default rates could be carried out by plaintiffs just as well as by defendants. Therefore, our theory would predict that lending discrimination should be regulated by a disparate treatment test, which of course it is not. In fact, to our knowledge differentials in default rates are not considered a valid defense in the enforcement of lending discrimination.

6.5 Housing discrimination

The situation in the housing discrimination arena is very similar to that in lending discrimination. There is agreement that a MDD type test should be applied, but there is disagreement as to what standard should be applied to the second leg of the test, whether a "business necessity" or a "legitimate business reason standard." The Third Circuit Court of Appeals in Resident Advisory Board v. Rizzo favors a business necessity standard.[12] But the Tenth Circuit rejected the "compelling business necessity" in favor of a "legitimate, non-pretextual justification." The Court stated that "the Secretary went beyond the business necessity test that the Supreme Court has enunciated in Title VII cases and incorrectly required that Mountain Side demonstrate a compelling need or necessity."[13] In most cases the defendant (developer, lender, etc.) probably has more information about intent to discriminate, and so a MDD-type burden-shifting test is predicted by our theory. But, as to how stringent a test should be applied in the second leg, the theory does not make sharp predictions. This is because housing discrimination cases are very heterogeneous with respect to ease of proving intent to discriminate. Proving intent may be more difficult in some housing discrimination cases (sales of real estate, zoning laws) and easier in others (for example, in the case of mortgage lending it might be easy to see the profitability of loans, which can be used to infer bias). Given this heterogeneity, the ambiguity observed in the case law is perhaps a reasonable posture.

7 Conclusions

The practical impact of anti-discrimination law depends crucially on what tests are applied to identify unlawful discrimination. Two tests, disparate impact and disparate treatment, represent the polar extremes. The burden-shifting (McDonnel-Douglas) test is in between. The conventional wisdom focuses on the "distributive" consequences of these tests (the first

[12] 564 F.2d 126, 149 (3d Cir. 1977). The court stated that the defendant's burden in Title VIII cases was to show that the policy or practice being challenged must serve, in theory and practice, a legitimate, bona fide interest of the Title VIII defendant, and the defendant must show that no alternative course of action could be adopted that would enable that interest to be served with less discriminatory impact.

[13] The Tenth Circuit rejected the "compelling business necessity standard advocated by HUD for private defendants. The court held that "the defendants had overcome the prima facie case by evidence of legitimate, non-pretextual justifications." The Court stated that "the Secretary went beyond the business necessity test that the Supreme Court has enunciated in Title VII cases and incorrectly required that Mountain Side demonstrate a compelling need or necessity." See Mountain Side Mobile estates Partnership v. Secretary of HUD, 56 F.3d 1243 (10th Cir. 1995).

is more plaintiff-friendly than the second). In this paper we have taken a different approach: we have derived these tests as the solution to the problem of a social planner who is trying to maximize the accuracy of the verdict minus the costs of producing evidence. We have then looked at several important areas of application of discrimination law, and we have tried to justify the different standards applied in each area on the basis of the predictions of our theory. Broadly speaking the theory has some predictive power in explaining the variation in doctrinal standards.

References

[1] Ian Ayres (2002), “Outcome Tests of Racial Disparities in Police Practices,” 4 J. JUST. RES. & STAT. ASSOC. 131 .

[2] Knowles, John, Persico, Nicola and Todd, Petra (2001) “Racial Bias in Motor Vehicle Searches: Theory and Evidence.” Journal of Political Economy 109(1), pp. 203-229.

[3] John R. Walter (1995) “The Fair Lending Laws and Their Enforcement.” Federal Reserve Bank of Richmond Economic Quarterly Volume 81/4 Fall 1995.

A Proofs

Proof of Proposition 1

Proof. $T_{1A}-T_{1W}$ is asymptotically distributed as a normal $N\left(\mu_B, \frac{\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{B_W}}{N_W}\right)$, where $\mu_B=\mu_{B_A}-\mu_{B_W}$. Therefore the probability of convicting the innocent is

$$\begin{aligned}
\Pr(CI) &= \Pr\left(T_{1A}-T_{1W}>t_1|\mu_B=0\right) \\
&= \Pr\left(\frac{T_{1A}-T_{1W}}{\sqrt{\frac{\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{B_W}}{N_W}}}>\frac{t_1}{\sqrt{\frac{\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{B_W}}{N_W}}}|\mu_B=0\right) \\
&= \Pr\left(N(0,1)>\frac{t_1}{\sqrt{\frac{\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{B_W}}{N_W}}}\right) \\
&= 1-\Phi\left(\frac{t_1}{\sqrt{\frac{\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{B_W}}{N_W}}}\right)
\end{aligned}$$

Similarly, the probability of acquitting the guilty is

$$\begin{aligned}
\Pr(AG) &= \Pr\left(T_{1A}-T_{1W}\leq t_1|\mu_B=K_B\right) \\
&= \Pr\left(\frac{T_{1A}-T_{1W}-K_B}{\sqrt{\frac{\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{B_W}}{N_W}}}\leq\frac{t_1-K_B}{\sqrt{\frac{\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{B_W}}{N_W}}}|\mu_B=K_B\right) \\
&= \Pr\left(N(0,1)\leq\frac{t_1-K_B}{\sqrt{\frac{\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{B_W}}{N_W}}}\right) \\
&= \Phi\left(\frac{t_1-K_B}{\sqrt{\frac{\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{B_W}}{N_W}}}\right)
\end{aligned}$$

Thus the loss function associated with t_1 is

$$L_1(y) = (1-\pi)\, l_{CI} \left[1-\Phi(y)\right] + \pi \; l_{AG} \; \Phi\left(y - \frac{K_B}{\sqrt{\frac{\sigma^2_{B_A}}{N_A} + \frac{\sigma^2_{B_W}}{N_W}}} \right), \tag{8}$$

where we have made the change of variables $y = \frac{t_1}{\sqrt{\frac{\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{B_W}}{N_W}}}$. This verifies the functional form of the loss function.

The derivative of the loss function with respect to y is

$$-(1-\pi)\, l_{CI} \phi(y) + \pi \; l_{AG} \; \phi(y-K) \tag{9}$$

where ϕ represents the p.d.f. of the standard Normal distribution. This function can be written as

$$\phi(y)\, \pi \; l_{AG} \left[\frac{\phi(y-K)}{\phi(y)} - \frac{(1-\pi)\, l_{CI}}{\pi \; l_{AG}} \right],$$

which shows that the derivative of the loss function has the same sign as the term in brackets. We want to show that the LHS of the term in brackets is increasing in y. It is increasing iff a monotone transformation is. Applying the log transformation yields

$$\begin{aligned} & \log(\phi(y-K)) - \log(\phi(y)) \\ = & -\frac{1}{2}\left[(y-K)^2 - y^2\right] \\ = & -\frac{1}{2}\left[K^2 - 2yK\right] \end{aligned}$$

which is increasing in y. This shows that the LHS is increasing in y. It follows that the loss function is concave. Interiority of the optimum follows from a well-known property of the Normal distribution which is that the likelihood ratio $\frac{\phi(y-K)}{\phi(y)}$ goes to infinity as $y \to \infty$ and to $-\infty$ when $y \to -\infty$. The interior optimum is the point y^* at which the term in brackets is zero; it follows that y^* is increasing in $\frac{(1-\pi) l_{CI}}{\pi \; l_{AG}}$. ■

Proof of Proposition 2

Proof. $T_{0W} - T_{0A}$ is asymptotically Normally distributed with mean $\mu_B - \mu_\Gamma$ and variance

$\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}$. Therefore the probability of convicting the innocent defendant is

$$\begin{aligned}
& \Pr\left(T_{0W}-T_{0A}>t_0|\mu_{B_A}-\mu_{B_W}=0\right) \\
= \; & \Pr\left(\frac{T_{0W}-T_{0A}+\mu_\Gamma}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}}>\frac{t_0+\mu_\Gamma}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}}|\mu_B=0\right) \\
= \; & \Pr\left(N\left(0,1\right)>\frac{t_0+\mu_\Gamma}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}}\right) \\
= \; & 1-\Phi\left(\frac{t_0+\mu_\Gamma}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}}\right)
\end{aligned}$$

The probability of acquitting the guilty is

$$\begin{aligned}
& \Pr\left(T_{0W}-T_{0A}<t_0|\mu_{B_A}-\mu_{B_W}=K_B\right) \\
= \; & \Pr\left(\frac{T_{0W}-T_{0A}+\mu_\Gamma-K_B}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}}<\frac{t_0+\mu_\Gamma-K_B}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}}|\mu_{B_A}-\mu_{B_W}=K_B\right) \\
= \; & \Pr\left(N\left(0,1\right)<\frac{t_0+\mu_\Gamma-K_B}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}}\right) \\
= \; & \Phi\left(\frac{t_0+\mu_\Gamma-K_B}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}}\right)
\end{aligned}$$

The expected loss from using this criterion is

$$\left(1-\pi\right)l_{CI}\left[1-\Phi\left(\frac{t_0+\mu_\Gamma}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}}\right)\right]+\pi\ l_{AG}\ \Phi\left(\frac{t_0+\mu_\Gamma-K_B}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A}+\frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}}\right)$$

The loss-minimizing t_0^* is obtained by minimizing this function. Denote

$$L_0(y) = (1-\pi)\, l_{CI}\, [1-\Phi(y)] + \pi\ l_{AG}\ \Phi\left(y - \frac{K_B}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A} + \frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}} \right),$$

where we have made the change of variables $y = \frac{t_0+\mu_\Gamma}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A} + \frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}}$. Then the loss-minimization problem can be stated as

$$\min_y L_0(y).$$

The solution y_0^* is linked to the optimal standard of proof t_0^* by the relationship $y_0^* = \frac{t_0^*+\mu_\Gamma}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A} + \frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}}$.

The rest of the properties are proved exactly as in the proof of Proposition 1. ■

Proof of Proposition 3

Proof. a. The planner chooses to verify abilities iff $L_1(y_1^*) + C(\gamma) \leq L_0(y_0^*)$. Rearranging this inequality yields

$$V(\gamma) \geq C(\gamma).$$

b. μ_Γ does not appear in either $L_0(y)$ or $L_1(y)$.

c. $V(\gamma)$ increases monotonically in $\frac{\sigma^2_{\Gamma_A}}{N_A} + \frac{\sigma^2_{\Gamma_W}}{N_W}$. Fix $\frac{\sigma^2_{\Gamma_A}}{N_A} + \frac{\sigma^2_{\Gamma_W}}{N_W}$ and compute $V(\gamma) = \min_y L_0(y) - \min_y L_1(y)$. Now increase $\frac{\sigma^2_{\Gamma_A}}{N_A} + \frac{\sigma^2_{\Gamma_W}}{N_W}$, thus giving rise to a loss function $\overline{L_0}(y) > L_0(y)$ for all y. The new value of information is $\min_y \overline{L_0}(y) - \min_y L_1(y) > \min_y L_0(y) - \min_y L_1(y) = V(\gamma)$.

d. First we show that $V(\gamma)$ is strictly greater than 0. Since for any given y

$$\Phi\left(y - \frac{K_B}{\sqrt{\frac{\sigma^2_{\Gamma_A}+\sigma^2_{B_A}}{N_A} + \frac{\sigma^2_{\Gamma_W}+\sigma^2_{B_W}}{N_W}}} \right) > \Phi\left(y - \frac{K_B}{\sqrt{\frac{\sigma^2_{B_A}}{N_A} + \frac{\sigma^2_{B_W}}{N_W}}} \right),$$

it follows that $L_1(y) < L_0(y)$ for all y. Therefore $\min_y L_1(y) < \min_y L_0(y)$ or, equivalently, $L_1(y_1^*) < L_0(y_0^*)$.

Next we show that $V(\gamma)$ is bounded above.

$$\begin{aligned}
V(\gamma) &= L_0\left(y_0^*\right) - L_1\left(y_1^*\right) < L_0\left(y_1^*\right) - L_1\left(y_1^*\right) \\
&= \pi\ l_{AG}\left[\Phi\left(y_1^* - \frac{K_B}{\sqrt{\frac{\sigma_{\Gamma_A}^2+\sigma_{B_A}^2}{N_A} + \frac{\sigma_{\Gamma_W}^2+\sigma_{B_W}^2}{N_W}}}\right) - \Phi\left(y_1^* - \frac{K_B}{\sqrt{\frac{\sigma_{B_A}^2}{N_A} + \frac{\sigma_{B_W}^2}{N_W}}}\right)\right] \\
&\leq \max_y \pi\ l_{AG}\left[\Phi\left(y - \frac{K_B}{\sqrt{\frac{\sigma_{\Gamma_A}^2+\sigma_{B_A}^2}{N_A} + \frac{\sigma_{\Gamma_W}^2+\sigma_{B_W}^2}{N_W}}}\right) - \Phi\left(y - \frac{K_B}{\sqrt{\frac{\sigma_{B_A}^2}{N_A} + \frac{\sigma_{B_W}^2}{N_W}}}\right)\right].
\end{aligned}$$

■

B The case when no information is optimal

If no evidence is acquired then the planner's loss is

$$L_\varnothing = \min\left\{(1-\pi)\, l_{CI}, \pi\ l_{AG}\right\}.$$

If the planner only collects evidence about $\{x_{iR}^*\}$ then the planner suffers

$$L_0\left(y_0^*\right) + C_B.$$

If the planner collects evidence about $\{\gamma_{iR}, x_{iR}^*\}$ then the planner suffers

$$L_1\left(y_1^*\right) + C_B + C_\gamma.$$

The planner who is considering whether to incur the cost of verifying abilities solves

$$\min\left\{L_0\left(y_0^*\right), L_1\left(y_1^*\right) + C_\gamma\right\}.$$

If this quantity is greater than $L_\varnothing - C_B$ then it is optimal for the planner to collect no evidence. We make the assumption that collecting some evidence, i.e., having some sort of discovery is efficient.

Assumption 6 $\min\left\{L_0\left(y_0^*\right), L_1\left(y_1^*\right) + C_\gamma\right\} < L_\varnothing - C_B$.

If this assumption is verified, the only question that remains is whether it is optimal for the planner to collect evidence on $\{x_{iR}^*\}$ only, or on $\{\gamma_{iR}, x_{iR}^*\}$.