A Representation Index: Glass Ceilings and Sticky Floors

Krishna Pendakur and Ravi Pendakur and Simon Woodcock

Simon Fraser University

June 2006

A Representation Index: Glass Ceilings and Sticky Floors

Krishna Pendakur
Economics, Simon Fraser University

Ravi Pendakur
Public and International Affairs, University of Ottawa

Simon Woodcock
Economics, Simon Fraser University

December 11, 2006

Abstract

Recent research on glass ceilings and sticky floors has focused on the magnitude of differences between groups in the upper and lower quantile cutoffs of the conditional wage distribution. However, quantile cutoffs for different groups are only weakly informative of representation. For example, if the top decile cutoff is lower for minority than majority workers, this tells us that minority workers are under-represented in the top decile, but does not tell us the magnitude of the under-representation. In this paper, we propose a direct measure of the representation of a population subgroup, which we define as the proportion of group members whose earnings lie below (or above) a population earnings quantile. Our representation index is easily generalized to condition on characteristics (such as age, education, etc). Further, it generalizes naturally to an index of the *severity* (or cost) of under-representation to group members, which is based on dollar-weighted representation. Both representation and severity indices are easily calculated via existing regression techniques. We illustrate the approach using Canadian earnings data.

JEL Codes: C1, C44, J71,

Keywords: representation, glass ceiling, discrimination, quantile regression, expectile regression

1 Introduction

It is well established that women and some ethnic minorities earn less than comparable white males (see e.g., Blau and Kahn (2000), Smith and Welch (1989), Pendakur and Pendakur (2002)). One proposed explanation is that workers in these groups face a 'glass ceiling' that limits their access to the very best jobs in society. Another possible explanation is the existence of a 'sticky floor' that crowds these workers into the very worst jobs in society. Both of these mechanisms suggest that disadvantaged groups will be under-represented in some parts of the earnings distribution and over-represented in others. In this paper, we present a new index to measure the representation of population subgroups in different parts of the population-level conditional earnings distribution. Our representation indices shed light on both the existence and consequences of glass ceilings and sticky floors.

Our representation index is a useful addition to the applied researcher's toolkit for at least three reasons. First, it provides an intuitive and direct measure of a group's prevalence in (or access to) a region of the population distribution of income (or earnings, wages, etc.) – in both the unconditional and conditional sense. That is, under-representation in the upper tail of the income distribution (for example) is an intuitive measure of disenfranchisement, and directly measures the presumed consequence of a glass ceiling. Second, our index is (in principle) completely nonparametric. It imposes no structure on the joint distribution of income and covariates. Third, it is easy to implement with standard statistical methods and popular software packages.

Although the idea of a glass ceiling is widespread, there has been surprisingly little research by economists to establish its existence or assess its consequences. Recent research has focused on the magnitude of differences between groups in the upper and lower quantile cutoffs of the conditional wage distribution (Fortin and Lemieux (1998), Albrecht et al. (2003), de la Rica et al. (2005), Kee (2005), Jellal et al. (2006)). For example, Albrecht et al. (2003) show in their study of Swedish data that the conditional top decile cutoff of women's wages is well below that of men and conclude that women face a glass ceiling. Pendakur and Pendakur (2006) use similar methods to study the earnings of ethnic minorities in Canada and find disparity in the upper and lower quantiles. However, knowing the location of a particular earnings quantile for different groups is only weakly informative of representation. Consider the case where the conditional top decile cutoff of earnings is \$10,000 lower for women than men. This tells us that women are under-represented in the top decile of the population conditional earnings distribution, but does not tell us the magnitude of the under-representation.

In this paper, we propose a different strategy to identify and measure the impact of a

glass ceiling or sticky floor. We define the *representation* of a population subgroup (hereafter 'group') as the proportion of group members whose earnings lie below (or above) a population earnings quantile. We say that a group is *under-represented* in a region of the earnings distribution if the proportion of the group in that region is smaller than the population proportion. Conversely, we say that a group is *over-represented* if the proportion of the group in that region is larger than the population proportion.

Consider a simple example. Suppose ten percent of the population earns more than $100,000. We say a group is under-represented in the top decile of the population earnings distribution if less than ten percent of the group earns more than $100,000. Of course a group's representation will depend in part on group members' characteristics (e.g., education, age, etc.), but we can easily generalize our notion of representation to condition on characteristics. We can estimate (for example, via quantile regression or nonparametric density estimation) the top decile cutoff of the population earnings distribution, conditional on observable characteristics, and ask what proportion of a particular group's members earn more than this characteristics-dependent cutoff. If the proportion is less than ten percent, we say that the group is under-represented in the top decile of the conditional earnings distribution.

Of course under-representation can take many forms. Disadvantaged group members may be clustered close to the quantile cutoff, or far below it. We therefore augment our representation measure with an index of the *severity* (or cost) of under-representation that weights representation by a function of dollar-distances from a cutoff.

Our approach has two advantages over comparison of quantile cutoffs to identify the presence of a glass ceiling or sticky floor. First, it provides a direct estimate of the *magnitude* of under-representation. Second, it provides a direct estimate of the *cost* of under-representation to the disadvantaged group.

There is a recent literature that seeks to decompose differences in the wage distribution between groups (or time periods) into differences due to characteristics and differences due to the returns to those characteristics. Examples include DiNardo et al. (1996), Donald et al. (2000), and Machado and Mata (2005), who base their decomposition on nonparametric reweighting, hazard function estimation, and quantile regression, respectively. In these papers, the object of interest is the entire marginal distribution of an outcome (e.g., wages). The primary application of the decomposition is to construct counterfactual distributions that would have prevailed if groups shared the same distribution of characteristics or returns to characteristics.

Our objective here is different: we are interested in measuring representation – a function of the joint distribution of outcomes and covariates – rather than estimating counterfactual

distributions of outcomes. Consequently, our representation index is complementary to these decomposition approaches in the following sense. One could construct a counterfactual distribution by one of these methods, and obtain a meaningful summary of the outcomes of group members that would have prevailed under the counterfactual distribution by comparing representation in the counterfactual to the realized distribution.

Furthermore, whereas the quantile cutoff approach and the decomposition approach both typically analyze the entire (marginal or conditional) distribution, our methods are informative and interesting even if only a single quantile is investigated. This is because the representation and severity indices summarize the entire (marginal or conditional) distribution above and below the quantile of interest. For example, a group's representation in the top decile of income is informative even if we know nothing about representation at any other quantile. This is particularly convenient when the researcher's interest centers on outcomes in a tail of the income distribution, as is the case when studying the consequences of a glass ceiling or sticky floor.

The paper proceeds as follows. We first develop a framework for modeling representation, considering both conditional and unconditional approaches. Then, we extend our framework to model the severity of under-representation. Finally, we illustrate the approach using Canadian data.

2 Modeling Representation

Consider the distribution of earnings given a vector of characteristics X. The population consists of individuals $i = 1, ..., N$ each of whom is member of a group $j = 1, ..., J$. Let N_j denote the number of members of group j. Let $f_j(y, X)$ represent the joint density of earnings and characteristics for group j, and denote the population joint density of earnings and characteristics by $f(y, X)$. The conditional cumulative distribution function (conditional cdf) of y given X for group j, $F_j(y|X)$, is defined as

$$F_j(y|X) = \frac{\int_0^y f_j(y, X)dy}{\int_0^\infty f_j(y, X)dy}. \tag{1}$$

The τ^{th} quantile of y conditional on X for group j, $q_j(\tau, X)$, is the inverse of this conditional cdf,

$$q_j(\tau, X) = F_j^{-1}(\cdot|X). \tag{2}$$

Here, we have that

$$F_j(q_j(\tau, X)|X) = \frac{\int_0^{q_j(\tau,X)} f_j(y, X)dy}{\int_0^{\infty} f_j(y, X)dy} = \tau.$$

Similarly, the τ^{th} quantile of y conditional on X for the population, $q(\tau, X)$, is defined by

$$F(q(\tau, X)|X) = \frac{\int_0^{q(\tau,X)} f(y, X)dy}{\int_0^{\infty} f(y, X)dy} = \tau \tag{3}$$

where $F(y|X)$ is the population conditional cdf.

We define the conditional representation function, $r_j(\tau, X)$, as the proportion of group j whose earnings lie below the τ^{th} population conditional quantile of y:

$$r_j(\tau, X) = F_j(q(\tau, X)|X). \tag{4}$$

The quantity $r_j(\tau, X)$ depends on X because the joint density of y and X may differ between the groups that comprise the population. If, for some value of X, this quantity exceeds τ, then at that value of X the group is over-represented in the region below τ; if it is less than τ, then the group is under-represented in that region.

When upper quantiles of the population conditional distribution of y given X are of primary interest, it may be more intuitive to compare $1 - r_j(\tau, X)$ to $1 - \tau$. In this way, we can compare representation *above* population quantiles rather than below. We will refer to such measures as 'above' measures, in contrast to the 'below' measures presented in this section. Typically, it will be convenient to use below measures to study sticky floors and above measures to study glass ceilings.

The conditional representation functions, $r_j(\tau, X)$ for $j = 1, ..., J$, in combination with the population conditional quantile function, $q(\tau, X)$, completely characterize the joint distribution of (y, X) for each group j. Thus, the set of functions $r_j(\tau, X)$ and $q(\tau, X)$ contain the same information as the set of quantile functions, $q_j(\tau, X)$.

Typically, quantile functions are not estimated for all possible values of τ; rather, they are estimated for a sparse grid of τ values, or sometimes even just a single value of τ. Thus, an important question is whether we learn more about glass ceilings and group representation in a region of the earnings distribution from the representation function or quantile function at a *single value* of τ. We argue that the representation function more directly illuminates the object of interest. Consider a simple example for a given vector of characteristics X. If the representation of minority workers in the top decile of earnings is 0.06, then they are under-represented by 40 percent. Alternately put, there are 40 percent fewer minority workers in the top decile of earnings than we would expect given their characteristics X. In contrast, if

we just use estimated quantiles and find that minority workers have a top decile cutoff that is \$10,000 below that of majority workers, we know that they are under-represented, but we don't know by how much. The representation function provides direct information on the object of interest: the degree to which a definable group of individuals is represented in a region of the conditional earnings distribution.[1]

The fact that $r_j(\tau, X)$ depends on X is desirable. It corresponds to a lack of parametric assumptions regarding the joint distribution of y and X in each group j. However, this lack of structure comes at a price. Evaluating the magnitude of $r_j(\tau, X)$ for any particular value of X is in principle a nonparametric problem that may have a very slow rate of convergence if X is high dimensional in its continuous elements (its discrete elements do not affect convergence rates). Furthermore, because $r_j(\tau, X)$ depends on X, its magnitude for any particular value of X is not revealing of representation for the group as a whole. A summary statistic based on averaging $r_j(\tau, X)$ solves both of these problems.

A convenient summary of representation for group j is the average value of $r_j(\tau, X)$ over all members of the group. By averaging over individuals, we implicitly average over their characteristics X with weights corresponding to the distribution of characteristics in group j. We denote this average as $r_j(\tau)$, and define our conditional representation index as

$$r_j(\tau) = \frac{1}{N_j} \sum_{i \epsilon j} r_j(\tau, X_i). \tag{5}$$

The conditional representation index, $r_j(\tau)$, is the average representation of group j below the τ^{th} population conditional quantile cutoff. If $r_j(\tau)$ exceeds τ, then the group as a whole is over-represented below the τ^{th} quantile of the population conditional distribution of y given X; if it is less than τ, then the group as a whole is under-represented in that region.

Since $r_j(\tau, X)$ is defined as a function of F and F_j (the joint cdfs of y and X of the population and group j) and because we place no restrictions on F and F_j, our conditional representation function and index are completely nonparametric. If the dimensionality of X is low, one can use nonparametric estimates of F and F_j to estimate representation. In cases where the dimensionality of X is too high for a completely nonparametric approach, the researcher can impose minimal parametric structure to facilitate estimation. One possibility would be to employ nonparametric single (or multiple) index density estimation. A quantile

[1] One of the important features of representation is that it captures an object of interest that is quite distinct from disparity in the conditional mean. For example, consider two groups whose joint distributions of y, X have finite support over y and which differ only in location of y by a factor d. Here, the disparity is the same at every quantile and all X, and the conditional mean functions also differ by d at all X. In this case, even with small d, there will be an upper tail of the population distribution where the disadvantaged group has zero representation.

regression-based approach is an easily implemented alternative, as we now demonstrate.

For any particular value of τ, we can easily estimate $r_j(\tau)$ using popular statistical software (such as R, S+, or Stata) in two steps. First, estimate the population conditional quantile cutoffs from the quantile regression of y on X. The population conditional quantile regression function, $Q(\tau, X)$, satisfies $P[y_i < Q(\tau, X)] = \tau$. We denote the quantile conditional *regression* function with Q rather than q because quantile regression typically imposes parametric structure on the problem, even though q is in essence a nonparametric object. Second, given an estimate of the population conditional quantile regression function $\hat{Q}(\tau, X)$, construct predicted values $\hat{Q}(\tau, X_i)$ for all i in group j. A sample estimate of the conditional representation index for group j is given by

$$\hat{r}_j(\tau) = \frac{1}{N_j} \sum_{i \epsilon j} I\left[y_i < \hat{Q}(\tau, X_i)\right]$$

where I is the indicator function. We note that $\hat{Q}(\tau, X_i)$ may not be unique in finite samples, but $\hat{r}_j(\tau)$ is unique.[2]

It is natural to ask under what circumstances the conditional representation function $r_j(\tau, X)$ is independent of X. That is, when does the conditional representation function $r_j(\tau, X)$ coincide with the conditional representation index $r_j(\tau)$? In fact, there may be little variation in $r_j(\tau, X)$ over X even if the conditional cdfs, $F_j(y|X)$, differ greatly across groups. This is because representation is invariant to some transformations of the joint distribution of y and X. The following proposition (proof in Appendix A) establishes that if the conditional quantile functions $q_j(\tau, X)$ have the same shape over X for all j, but possibly describe different quantiles for each j, then there is *no* variation in $r_j(\tau, X)$ over X.

Proposition 1 *If ϕ_j is a monotone increasing mapping from $[0,1]$ to $[0,1]$ and*

$$F_j(y|X) = \phi_j\left(F_1(y|X)\right) \tag{6}$$

for all $j = 2, ..., J$, then $r_j(\tau, X)$ is independent of X.

3 Accounting for Characteristics

In many instances, it will be of interest to assess how individual characteristics contribute to a group's over- or under-representation in a region of the earnings distribution. This is

[2] When the empirical cdf of $y|X_i$ has flat regions, quantile cutoffs in those regions are bounded but not unique. Because $\hat{r}_j(\tau)$ implicitly integrates over flat regions of the empirical cdf, $\hat{r}_j(\tau)$ is unique.

easily done by comparing the estimated conditional representation index for group j, $\hat{r}_j(\tau)$, to the estimated unconditional representation of group j in the population distribution of earnings.

The τ^{th} unconditional quantile cutoff of the population distribution, $q(\tau)$, solves

$$P[y_i < q(\tau)] = \tau.$$

Let $\hat{q}(\tau)$ denote a sample estimate of this quantity. Note that since $q(\tau)$ does not depend on X, neither quantile regression nor parametric structure need be employed to obtain the estimate $\hat{q}(\tau)$. Instead, we may obtain $\hat{q}(\tau)$ by sorting the data by Y. Our unconditional representation index for group j, $R_j(\tau)$, is the proportion of group members whose value of y_i lies below the τ^{th} unconditional population quantile cutoff: $R_j(\tau) = F_j\left(q\left(\tau\right)\right)$, where $F_j\left(y\right)$ is the unconditional cdf of earnings in group j. A sample estimate of the unconditional representation index for group j is

$$\hat{R}_j(\tau) = \frac{1}{N_j} \sum_{i \epsilon j} I\left[y_i < \hat{q}(\tau)\right], \tag{7}$$

i.e., the sample proportion of group j whose earnings lie below $\hat{q}\left(\tau\right)$.

Consider the representation of group j in the bottom decile of the earnings distribution. The value of $\hat{r}_j(0.10)$ gives the proportion of the group's members whose earnings are below the population cutoff for the conditional bottom decile, controlling for variation in X. In contrast, the value of $\hat{R}_j(0.10)$ gives the proportion of the group's members whose earnings are below the population cutoff for the bottom decile of the unconditional distribution, without controlling for variation in X. If, for example, $\hat{R}_j(0.10) = 0.20$ and $\hat{r}_j(0.10) = 0.10$, then we say that poor characteristics explain the over-representation of group j in the bottom decile of earnings. If, on the other hand, $\hat{R}_j(0.10) = 0.20$ and $\hat{r}_j(0.10) = 0.15$, we say that poor characteristics do not explain all of the over-representation of group j in the bottom decile, and that we observe 50% more members of group j in the bottom decile than in the population as a whole, even after controlling for their characteristics.

4 Severity of Under-Representation

It is natural to ask whether under-representation in a region of the earnings distribution has large or small consequences. If, for example, minorities are under-represented in the top decile but over-represented in the top percentile, then the representation index $r_j(0.90)$ would be above 0.90, but minority workers could actually be receiving much of the total

earnings in the upper tail of the distribution. In this case, high representation in the top percentile of the distribution might mitigate low representation in the top decile.

If we evaluated the representation index at all possible τ, we could examine the complete collection of representation indices. More realistically, however, if we only evaluate the representation index for a sparse grid on τ, or even a single value of τ, then it is desirable to have some aggregator to supplement the representation index. A natural aggregator of representation weights the representation below (above) a cutoff by (some function of) the dollar amount of the deviation below (above) the cutoff.

The *expectile* function (see Newey and Powell (1987)) defines a natural way to weight these deviations. It can be expressed as the solution to a weighted version of the quantile problem. The population quantile function given by (3) defines a cutoff q such that the proportion of the density of earnings below q is τ. In contrast, the expectile function defines a cutoff $e(\tau, X)$ such that the proportion of the *weighted* density of earnings below $e(\tau, X)$ is τ. The weight used in the expectile function is the dollar value of the distance from the cutoff. The expectile function, e, is thus defined by

$$\frac{\int_0^{e(\tau,X)} |e(\tau, X) - y| \ f(y, X)dy}{\int_0^{\infty} |e(\tau, X) - y| \ f(y, X)dy} = \tau, \tag{8}$$

which simply adds the weight $|e(\tau, X)-y|$ to (3). Unlike quantiles, expectiles are unique even if the cdf has flat regions. For people earnings less than the cutoff, $|e(\tau, X)-y| = e(\tau, X)-y$ gives the 'shortfall' of earnings below the cutoff, and for those earning more than the cutoff, $|e(\tau, X) - y| = y - e(\tau, X)$ gives the 'surplus' of earnings above the cutoff. The expectile function defines the cutoff value such that the total shortfall is a proportion τ of the total shortfall plus the total surplus. For $\tau = 0.50$, the total shortfall equals the total surplus, which characterizes the mean. Thus the expectile for $\tau = 0.50$ is the conditional mean, which may be estimated via ordinary least squares. Expectiles for other values of τ may be estimated via a type of weighted ordinary least squares.

We define the conditional severity function, $s_j(\tau, X)$, as the weighted representation below the population expectile $e(\tau, X)$, where the weight is the distance $|e(\tau, X) - y|$. That is,

$$s_j(\tau, X) = \frac{\int_0^{e(\tau,X)} |e(\tau, X) - y| \ f_j(y, X)dy}{\int_0^{\infty} |e(\tau, X) - y| \ f_j(y, X)dy}. \tag{9}$$

Note that $s_j(\tau, X) = \tau$ for all τ if and only if $f_j(y, X) = f(y, X)$, i.e., if group j has the same joint distribution of (y, X) as the population as a whole.

We argue that the conditional severity function captures the economic cost, or severity, of under-representation in two related ways. First, note that the weights increase with

distance from the cutoff $e(\tau, X)$. Therefore density far below the cutoff, where the cost of under-representation is greatest, is given greater weight than density just below the cutoff. If we hold density below the cutoff (representation) constant, the conditional severity function increases as the earnings distribution becomes more concentrated at very low levels of earnings, and decreases as the distribution becomes more concentrated just below the cutoff. In contrast, if we hold density above the cutoff constant, the conditional severity function decreases as the earnings distribution becomes more concentrated at very high levels of earnings, because such concentration increases the denominator of the severity function. Second, the conditional severity function can be interpreted in terms of conditional means of y. The numerator of (9) is a scaled difference between the expectile cutoff $e(\tau, X)$ and the conditional mean of y given X below the cutoff:

$$\begin{aligned} \frac{\int_0^{e(\tau,X)} |e(\tau,X) - y| \; f_j(y,X)dy}{\int_0^{e(\tau,X)} \; f_j(y,X)dy} &= e(\tau,X) - \frac{\int_0^{e(\tau,X)} y \; f_j(y,X)dy}{\int_0^{e(\tau,X)} \; f_j(y,X)dy} \\ &= e(\tau,X) - \mathbb{E}_j[y|X, y < e(\tau,X)] \end{aligned}$$

where $\mathbb{E}_j$ denotes the expectation for members of group j. Thus the numerator of (9) is large if the group's conditional mean of earnings below the cutoff is small. The denominator of (9) simply normalizes the conditional severity function to lie in $[0, 1]$.

Thus our severity measure has a natural metric. For a given X, the severity measure evaluated on the population as a whole equals τ by the definition of the population expectile. If $s_j(\tau, X)$ exceeds τ, then the dollar-weighted representation of group j below the τ^{th} population expectile exceeds the dollar-weighted representation of the population. If $s_j(\tau, X)$ is less than τ, then the dollar-weighted representation of group j below the τ^{th} population expectile is less than the dollar-weighted representation of the population.

The conditional severity function usefully supplements the conditional representation function. For example, if for a given X, representation for group j in the bottom population decile is 0.20, then the proportion of group j's members in the bottom decile of the distribution is twice that of the population as a whole. However, if the earnings of members of group j are clustered just below the bottom decile cutoff, then this over-representation in the lowest decile is not very severe. The conditional severity function would illuminate such a pattern. In this example, it might be the case that the severity measure for the 10th population expectile is 0.15, which would indicate that when weighted by dollars, the over-representation in the bottom of the distribution is not as severe as the representation index might suggest.

In the example above, we considered representation and severity with $\tau = 0.10$. Note,

however, that in general the dollar value of the τ^{th} population quantile will not be the same as the dollar value of the τ^{th} population expectile. We define our severity measure based on the expectile function to give it the natural metric described above. One could alternately define a conditional severity function directly from the population conditional quantile function, for example, as the dollar-weighted representation below $q(\tau, X)$. However, a conditional severity function defined in this way would not have a natural metric. In particular, its value for a group would only be meaningful relative to its value for the population as a whole. In addition, such a measure of conditional severity would not be unique, because $q(\tau, X)$ is not unique and hence neither are distances from $q(\tau, X)$.

Like the representation function, the severity function, $s_j(\tau, X)$, depends on X. We therefore desire a summary measure of severity for group j that averages over X, and define a conditional severity index, $s_j(\tau)$, as

$$s_j(\tau) = \frac{1}{N_j} \sum_{i \epsilon j} s_j(\tau, X_i). \tag{10}$$

Here, $s_j(\tau)$ is the average conditional severity for members of group j below the τ^{th} population expectile cutoff. If $s_j(\tau)$ is greater than τ, then the earnings of the group are crowded below the τ^{th} population expectile.

Replacing population quantities with sample estimates in (9) and (10) defines a sample estimate of the conditional severity index, $\hat{s}_j(\tau)$, based on a population-level estimated expectile function. The expectile function $e(\tau, X)$ is easily estimated by expectile regression, which is related to both ordinary least squares and quantile regression (see Newey and Powell (1987), especially footnote 2). The differences between these regression methods are most easily understood as a difference between the penalty function applied to deviations of y_i from a function, $f(\theta, X_i)$, which depends on parameters θ and covariates X.[3] Defining residuals $u_i = y_i - f(\theta, X_i)$, all three regression approaches minimize (by choice of θ) the sum of penalized residuals, $\sum_{i=1}^{N} p(u_i)$. In ordinary least squares, the penalty function is the square function, $p(u) = u^2$. In quantile regression, the penalty function is the weighted absolute value function, $p(u) = |\tau - I(u < 0)| \cdot |u|$. In expectile regression, the penalty function is the weighted square function, $p(u) = |\tau - I(u < 0)| \cdot u^2$. Thus, expectile regression combines features of quantile regression and mean (OLS) regression.

For any value of τ, the severity index is easily estimated in two steps. The first step is to estimate the population expectile regression function $\hat{E}(\tau, X)$. Again, we use the notation

[3] For expositional clarity we describe the regression functions as parametric. In principle, we could replace the parametric regression function $f(\theta, X_i)$ with a nonparametric regression function for all three types of regression.

$\hat{E}$ rather $\hat{e}$ because expectile *regression* typically embodies parametric structure even though the expectile is defined without structure on F or $F_{j.}$ Expectile regressions may be estimated via iterated asymmetrically weighted least squares (see Newey and Powell (1987)) as follows: given a pre-estimate of the regression function $f(\theta, X_i)$, compute weights $|\tau - I(u_i < 0)|$ and estimate the regression of y on X by weighted least squares (WLS). Then, update the weights using the new estimates, and re-estimate the model by WLS. This process is repeated to convergence, and the resulting regression model is the estimate $\hat{E}(\tau, X)$.

Given an estimate $\hat{E}(\tau, X)$ of the population expectile regression function, the second step is to construct predicted values $\hat{E}(\tau, X_i)$ for all i in group j, and compute $\hat{s}_j(\tau)$ as the sample average of weighted representation below $\hat{E}(\tau, X_i)$:

$$\hat{s}_j(\tau) = \frac{\sum_{i \epsilon j} \max\left\{\hat{E}(\tau, X_i) - y_i, 0\right\}}{\sum_{i \epsilon j} |\hat{E}(\tau, X_i) - y_i|}.$$

We define an unconditional severity index, $S_j(\tau)$, analogously to the unconditional representation index $R_j(\tau)$. A sample estimate of the unconditional τ^{th} expectile cutoff of the population distribution, $e(\tau)$, solves

$$\frac{\sum_{i=1}^{N} \max\left\{\hat{e}(\tau) - y_i, 0\right\}}{\sum_{i=1}^{N} |\hat{e}(\tau) - y_i|} = \tau$$

for $\hat{e}(\tau)$. Again, since $e(\tau)$ does not depend on X, neither expectile regression nor parametric structure need be employed to obtain the estimate $\hat{e}(\tau)$. Instead, we may obtain $\hat{e}(\tau)$ by sorting the data by Y, and solving for $\hat{e}(\tau)$. A sample estimate of the unconditional severity index for group j is

$$\hat{S}_j(\tau) = \frac{\sum_{i \epsilon j} \max\left\{\hat{e}(\tau) - y_i, 0\right\}}{\sum_{i \epsilon j} |\hat{e}(\tau) - y_i|}.$$

As in the case of the representation index, we can compare the conditional severity index to the unconditional severity index to assess the contribution of individual characteristics to the severity of under-representation.

We close our discussion of the severity function and index by noting that its value and interpretation depend on the scale of the outcome variable y. This is in contrast to representation, which is invariant to transformations of y. We point this out because it is typical in applied labour economics to estimate models where the dependent variable is measured in logarithms. Indeed, this is the case in our application below. This implies that the weights should be interpreted as log-dollar distances, or approximately as percentage distances, from the population expectile cutoff.

5 Results

We estimate our representation and severity indices on the universe of long form responses to the 2001 Census of Canada. These are confidential data, so we discuss replicability below and in Appendix B. Census long forms are administered to twenty percent of Canadian households, with the exception of households on Aboriginal reserves that are sampled at a 100 percent rate. All reported estimates are computed using sample weights provided by Statistics Canada.[4]

We define three broad ethnic categories of interest: Aboriginal, visible minority and white. These categories correspond to those used in Canadian Employment Equity policy. A person is classified as Aboriginal if their self-reported ancestry includes Aboriginal, Métis, Inuit, or North American Indian. A person is classified as visible minority if they are not Aboriginal, and their self-reported ancestry includes any region other than Canada, the United States, Europe, Israel, Australia or New Zealand. All others are classified as white.

Our focus is on the native-born population, and our primary interest is on non-white ethnic minorities. We focus on the native-born population to eliminate the potentially confounding effects of immigration on the earnings distribution. Canadian-born visible minorities comprise less than 2 percent of the Canadian-born population, and Aboriginals comprise less than 3 percent. Estimation and inference therefore requires a large sample, so the universe of long form Census responses is ideally suited to this investigation. Our analysis sample consists of all Canadian-born residents of Canada, 25 to 64 years of age, whose primary source of income is from wages and salaries, and who report positive schooling and earnings.

We base the representation and severity indices on a frequently used measure of labor earnings: the natural logarithm of annual gross earnings from wages and salaries. The conditional indices control for age (8 categories), schooling (13 categories), marital status (5 categories), household size, official language knowledge (3 categories), and 12 area-of-residence categories comprised of 10 Census Metro Areas (CMAs), a small CMA identifier, and a non-CMA identifier.

Although Statistics Canada guidelines do not allow release of the exact counts of population groups in our analysis, our analysis sample contains approximately 900,000 observations each for men and women. In the interest of replicability, we present estimates based on the Public Use Microdata File (PUMF) of the 2001 Census of Canada in the appendix. We do not report estimates from the PUMF in the body of the paper because the PUMF has far

[4]Sample weights are constructed to replicate population counts by age, sex, marital status, mother tongue, and household composition. See Statistics Canada (2003) for details.

fewer observations than the (confidential) database that we use. Appendix Table 1 reports sample means in the PUMF, subject to the sample restrictions defined above. Weighted sample means in our analysis sample match those in Appendix Table 1 to at least two decimal places. The sample statistics contain no surprises. There is considerable dispersion in earnings across demographic groups: the average earnings of men exceed those of women, and the average earnings of whites exceed those of visible minorities and Aboriginal persons.

Table 1 presents estimates of the conditional and unconditional representation index at the tenth, fiftieth, and ninetieth percentile of log earnings. At each of these quantiles, the representation of white men and women corresponds very closely to that of the entire native-born population. This is unsurprising, given that white men and women comprise over 95 percent of the native-born population. However, Aboriginals and visible minorities are heavily over-represented below the tenth and fiftieth percentiles, and under-represented above the ninetieth percentile. In general, the magnitude of the representation index is more extreme for Aboriginals than visible minorities, and interestingly, is more extreme for men than for women. Recall that the indices in this paper are all presented as 'below' measures, but can be characterized as 'above' measures by subtracting them from 1. In our discussion below, we will focus mainly on the top and bottom deciles, and will discuss bottom decile results with below measures in mind, and upper decile results with above measures in mind.

We begin a closer inspection of Table 1 with the least extreme group, female visible minorities. Compared to the population of women, visible minority women are unconditionally over-represented by almost 50 percent in the bottom decile of log earnings ($\hat{R}_j(0.10) = 0.149$), and under-represented by nearly 20 percent in the top decile ($\hat{R}_j(0.90) = 0.919$). However, these values are almost completely explained by the characteristics of group members ($\hat{r}_j(0.10) = 0.104, \hat{r}_j(0.90) = 0.904$).

Among men, visible minorities are quite heavily over-represented in the lower tail of the distribution and under-represented in the upper tail: unconditionally, there are fully 2.26 times more male visible minorities below the tenth percentile of log earnings ($\hat{R}_j(0.10) = 0.226$), and 41 percent fewer above the ninetieth percentile, than in the population ($\hat{R}_j(0.90) = 0.941$). This is largely, but not completely, explained by their characteristics. Controlling for individual characteristics reduces the representation index at the tenth percentile to 0.129, and at the ninetieth percentile to 0.924. These results suggest that male visible minorities face not only a glass ceiling that limits their opportunities at the top of the earnings distribution, but also a sticky floor that limits their advancement beyond the lowest-paying jobs in society.

Among both men and women, Aboriginals fare worse than visible minorities. Unconditionally, Aboriginal women are over-represented by 86 percent in the bottom decile of log

earnings and under-represented by 58 percent in the top decile. The situation is even bleaker for Aboriginal men, more than 70 percent of whom earn less than the median log earnings of all native-born men. They are over-represented by 119 percent in the bottom decile and under-represented by 66 percent in the top decile. Accounting for characteristics explains about half of the disparity for women, but little of the disparity for men. Controlling for individual characteristics, Aboriginal women remain over-represented in the bottom decile by 42 percent, and under-represented in the top decile by 18 percent. Likewise, Aboriginal men remain over-represented by 102 percent in the bottom decile, and under-represented by 33 percent in the top decile. Even after controlling for characteristics, nearly 66 percent of Aboriginal men earn less than the median. We take these results as strong evidence that Aboriginal men and women face a substantial glass ceiling, and an even more substantial sticky floor.

Table 2 presents estimates of the severity index. For most groups, they tell a qualitatively similar story to that of the representation index. However, we see that the representation index substantially understates the impact of the glass ceiling and sticky floor for Aboriginal men. For this group, the severity index at the mean, $\hat{S}_j(0.50)$, is 0.856. Weighted by (log-earnings) distance from the mean, the representation of Aboriginal men below the 0.50 expectile – the mean – is far higher than the population of native-born men. Indeed, weighting by distance paints a more dismal picture than considering (unweighted) representation below the median of log earnings, $\widehat{R}_j(0.50) = 0.705$. This is because the earnings of Aboriginal men are concentrated in the lowest part of the log earnings distribution.

The severity of over-representation below the mean is substantially reduced when one accounts for the characteristics of Aboriginal males. The conditional severity index at the mean, $\widehat{s}_j(0.50)$, is 0.731, which is 0.125 lower than the unconditional severity index. This remains very large, however, as we can see from the corresponding above measure. Even controlling for characteristics leaves the weighted representation (that is, severity) of Aboriginal men above the conditional mean at approximately half of that of all native-born men.

6 Conclusion

The representation index provides an intuitive and easily computed measure of a group's representation in a region of the earnings distribution. The index may be formulated to condition on observable characteristics, or not. We augment the representation index with a severity index that weights representation by the distance from a cutoff, and so provides a measure of the economic cost, or severity, of under-representation to the under-represented group. In conjunction, the representation and severity indices provide a comprehensive

picture of under- and over-representation and its economic consequences. They represent an important addition to the toolkit of applied researchers studying glass ceiling and sticky floor phenomena.

In our application to Canadian data, we find strong evidence that Aboriginals and visible minorities are under-represented in the conditional upper decile of the population earnings distribution, and are over-represented in the conditional lower decile of the population earnings distribution. This suggests that these groups face both glass ceilings and sticky floors.

A Appendix: Omitted Proofs

Proof of Proposition 1. In equation (6), ϕ_j maps the conditional cumulative distribution of y given X from group to group. Since the quantile function, q_j, is the inverse of F_j, an implication of (6) is

$$F_j(q_1(\tau, X)|X) = \phi_j(\tau)$$

which is independent of X. If, for example, $\phi_2(0.5) = 0.6$, this implies that the median earnings for group 1 is the 60th percentile of earnings for group 2 at all values of X. The restriction (6) ensures that the τ^{th} quantile of y conditional on X for the population has the same shape as *some* quantile for any group in the population.

Let $\pi_j \equiv N_j/N$ denote the proportion of the population that belongs to group j, so that $\sum_{j=1}^{J} \pi_j = 1$. Given the restriction (6), the population-level quantiles are implicitly defined by

$$\sum_{j=1}^{J} \pi_j \phi_j \left(F_1(q(\tau, X)|X)\right) = \tau.$$

Since the left-hand side is a weighted sum of monotone increasing functions of a single argument $F_1(q(\tau, X)|X)$, it is invertible with respect to this argument, and there must exist a monotone increasing function ϕ_1 such that

$$F_1(q(\tau, X)|X) = \phi_1(\tau).$$

Here, ϕ_1 is similar to ϕ_j for $j = 2, ..., J$, in that it is independent of X, but differs in that ϕ_1 maps from the population-level quantiles into the conditional cdf of group 1.

The representation of group j is then given by

$$F_j(q(\tau, X)|X) = \phi_j(\phi_1(\tau))$$

and it is independent of X. ■

B Appendix: Replicability

In the interest of replicability, we estimate the representation and severity indices on the Public Use Microdata File (PUMF) of the 2001 Census of Canada. These are presented in Appendix Tables 2 and 3. The conditional measures correspond very closely to those obtained on the universe of long form responses (Tables 1 and 2). However, there are notable discrepancies between the unconditional representation and severity estimates in the PUMF and the universe data. This is to be expected, given the nature of the sample weights in the two files (all our estimates are computed using sample weights provided by Statistics Canada). In particular, the sample weights are designed to match population counts by age, sex, marital status, mother tongue, and household composition (see Statistics Canada (2003) for details). However, they do not directly depend on the distribution of earnings. Thus we observe significant differences in the unconditional estimates, but this difference vanishes when we condition on age, sex, marital status, mother tongue, and household composition.[5]

References

Albrecht, J., A. Bjorklund, and S. Vroman (2003). Is there a glass ceiling in Sweden? *Journal of Labor Economics 21*(1), 145–177.

Blau, F. D. and L. M. Kahn (2000). Gender differences in pay. *Journal of Economic Perspectives 14*(4), 75–99.

de la Rica, S., J. J. Dolado, and V. Llorens (2005). Ceilings and floors: Gender wage gaps by education in Spain. IZA Discussion Paper No. 1483.

DiNardo, J., N. Fortin, and T. Lemieux (1996). Labour market institutions and the distribution of wages. *Econometrica 64*, 1001–1044.

Donald, S., D. Green, and H. Paarsch (2000). Differences in the wage distributions between Canada and the United States: An application of a flexible estimator of the distribution function in the presence of covariates. *Review of Economic Studies 67*, 609–633.

Fortin, N. M. and T. Lemieux (1998). Rank regressions, wage distributions, and the gender gap. *The Journal of Human Resources 33*(3), 610–643.

Jellal, M., C. Nordman, and F.-C. Wolff (2006). Theory and evidence on the glass ceiling effect using matched worker-firm data. Document de Travail DIAL DT/2006-3.

[5] As described in section 5, the conditional estimates also control for educational attainment and region of residence.

Kee, H. J. (2005). Glass ceiling or sticky floor? Exploring the Australian gender pay gap using quantile regression and counterfactual decomposition methods. The Australian National University Center for Economic Policy Research Discussion Paper No. 487.

Machado, J. A. F. and J. Mata (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics 20*, 445–465.

Newey, W. K. and J. L. Powell (1987). Asymmetric least squares estimation and testing. *Econometrica 55*(4), 819–847.

Pendakur, K. and R. Pendakur (2002). Colour my world: Has the minority-majority earnings gaps changed over time? *Canadian Public Policy 28*(4), 489–512.

Pendakur, K. and R. Pendakur (2006). Glass ceilings for ethnic minorities.

Smith, J. P. and F. R. Welch (1989). Black economic progress after Myrdal. *Journal of Economic Literature XXVII*, 519–564.

Statistics Canada (2003). *2001 Census Handbook*. Statistics Canada, Catalogue No. 92-379-XIE.

Table 1
Representation Index for Selected Demographic Groups

	Unconditional			Conditional		
	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
Women						
White	0.098	0.494	0.901	0.098	0.497	0.897
Visible Minorities	0.149	0.559	0.919	0.104	0.507	0.904
Aboriginal Persons	0.186	0.643	0.958	0.142	0.560	0.918
Men						
White	0.099	0.489	0.898	0.096	0.493	0.896
Visible Minorities	0.226	0.672	0.941	0.129	0.555	0.924
Aboriginal Persons	0.219	0.705	0.966	0.202	0.656	0.933

Source: Author's calculations based on all long form responses to the 2001 Census of Canada. Simulated standard errors are available on request. All standard errors are less than 0.002.

Table 2
Severity Index for Selected Demographic Groups

	Unconditional			Conditional		
	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
Women						
White	0.095	0.488	0.895	0.098	0.495	0.898
Visible Minorities	0.153	0.615	0.929	0.117	0.531	0.913
Aboriginal Persons	0.242	0.747	0.971	0.156	0.615	0.932
Men						
White	0.093	0.479	0.893	0.096	0.488	0.896
Visible Minorities	0.231	0.743	0.958	0.138	0.594	0.936
Aboriginal Persons	0.334	0.856	0.986	0.224	0.731	0.957

Source: Author's calculations based on all long form responses to the 2001 Census of Canada. Simulated standard errors are available on request. All standard errors are less than 0.001.

Appendix Table 1
Summary Statistics in the Public Use Microdata File (PUMF)

	Men		Women	
	Mean	Std. Dev.	Mean	Std. Dev.
ln(Earnings)				
White	10.4	0.96	9.92	1.07
Visible Minorities	10.3	1.08	10.0	1.15
Aboriginal Persons	9.86	1.22	9.52	1.27
Age (years)	41.2	9.86	41.0	9.64
Number of household members	3.01	1.33	2.98	1.29
Single-person household (percent in category)	12.2		11.0	
Ethnicity (column percent in category)				
White	95.5		95.4	
Visible Minorities	1.63		1.65	
Aboriginal Persons	2.90		2.91	
Knowledge of Official Languages (column percent in category)				
English only	64.5		63.9	
French only	12.5		13.8	
Both English and French	23.0		22.3	
Highest level of educational attainment (column percent in category)				
Less than grade 5	0.50		0.28	
Ggrades 5 to 8	3.22		1.87	
Grades 9 to 13	16.2		12.7	
High school graduate	14.1		16.0	
Trades certificate or diploma	5.42		2.95	
College, without college or trades certificate or diploma	6.40		6.75	
College, with trades certificate or diploma	11.8		6.20	
College, with college certificate or diploma	13.8		20.3	
University, without college certificate, diploma, or degree	3.59		3.21	
University, with certicate/diploma below bachelor level	6.38		8.38	
University, with bachelor or first professional degree	12.9		15.3	
University, with university certificate above bachelor level	1.70		2.54	
University, with master's degree[s]	3.51		3.14	
University, with earned doctorate	0.63		0.32	
Marital Status (column percent in category)				
Single. never married	20.3		16.3	
Married, including common-law	71.5		70.1	
Separated	2.68		3.79	
Divorced	5.10		8.24	
Widowed	0.39		1.65	
Region of Residence (column percent in category)				
Montreal	11.9		12.2	
Toronto	10.2		10.6	
Vancouver	5.13		5.16	
All other Census Metropolitan Areas (CMAs)	31.7		31.5	
Not in a CMA	41.1		40.5	
Number of Observations	118,203		114,682	

Source: Author's calculations based on the Public Use Microdata File (PUMF) of the 2001 Census of Canada.

Appendix Table 2
Representation Index for Selected Demographic Groups, PUMF

	Unconditional			Conditional		
	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
Women						
White	0.098	0.501	0.899	0.098	0.497	0.900
Visible Minorities	0.093	0.450	0.883	0.107	0.522	0.898
Aboriginal Persons	0.179	0.668	0.954	0.140	0.562	0.908
Men						
White	0.098	0.511	0.898	0.096	0.494	0.898
Visible Minorities	0.145	0.592	0.910	0.126	0.562	0.930
Aboriginal Persons	0.254	0.732	0.964	0.202	0.664	0.928

Source: Author's calculations based on the Public Use Microdata File (PUMF) of the 2001 Census of Canada. Simulated standard errors are available on request. All standard errors are less than 0.006.

Appendix Table 3
Severity Index for Selected Demographic Groups, PUMF

	Unconditional			Conditional		
	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
Women						
White	0.097	0.493	0.898	0.098	0.496	0.899
Visible Minorities	0.102	0.454	0.873	0.124	0.535	0.904
Aboriginal Persons	0.212	0.725	0.964	0.142	0.594	0.922
Men						
White	0.094	0.485	0.895	0.095	0.488	0.897
Visible Minorities	0.146	0.609	0.927	0.142	0.602	0.937
Aboriginal Persons	0.292	0.813	0.979	0.232	0.739	0.956

Source: Author's calculations based on the Public Use Microdata File (PUMF) of the 2001 Census of Canada. Simulated standard errors are available on request. All standard errors are less than 0.003.