A Representation Index: Measuring the Representation of Minorities in the Income Distribution

Krishna Pendakur
Economics, Simon Fraser University

Ravi Pendakur
Public and International Affairs, University of Ottawa

Simon Woodcock
Economics, Simon Fraser University

August 21, 2008

Abstract

The existence of glass ceilings and sticky floors suggests that disadvantaged workers will be under-represented in some parts of the income distribution, and over-represented in others. We present a *representation index* that measures the prevalence of population subgroups in different regions of the income (or any other) distribution. 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 Census data on the earnings of ethnic minorities.

JEL Codes: C1, C44, J71

Keywords: representation, discrimination, glass ceiling, 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 these workers face a "glass ceiling" that limits their access to society's best jobs. Another possible explanation is the existence of a "sticky floor" that crowds these workers into the worst jobs. Both 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 regions of the income distribution. Our representation index sheds light on the existence and consequences of glass ceilings and sticky floors, and on minority wage outcomes more generally.

Our representation index is a useful addition to the applied researcher's toolkit. It provides an intuitive and direct measure of a group's prevalence in (or access to) a region of the income distribution – in both conditional and unconditional senses. Our index is also (in principle) completely nonparametric. It imposes no structure on the joint distribution of income and covariates. Furthermore, it is easy to implement with standard statistical methods and popular software packages.

Recent research has focused on the magnitude of wage differentials in upper and lower quantiles of the conditional wage distribution (e.g., Fortin and Lemieux (1998), Albrecht et al. (2003)). For example, Albrecht et al. (2003) find evidence of a glass ceiling in Sweden based on the male-female wage differential at various quantiles. Pendakur and Pendakur (2007) 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 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 their under-representation.

In this paper, we propose and define a *representation index* which directly measures a group's representation in a region of the income (or any other ordered) distribution. We define the representation of a population subgroup (hereafter "group") as the proportion of group members whose income lies below (or above) the τ^{th} income quantile of an anchoring distribution. The anchoring distribution can be that of the whole population, or of another population subgroup (e.g., majority workers) and can condition on observable covariates. We say that a group is under-represented in a region of the income

distribution if the proportion of the group's members in that region is less than τ. Conversely, we say that a group is over-represented if the proportion of the group in that region is larger than τ.

Our index is intuitive, and indeed similar measures have been casually used in the literatures on income distribution and education. For example, Kopczuk et al. (2007) use a similar measure to assess the representation of women in the upper part of the income distribution. Our contribution is to formalize and synthesize these measures, and to show how they may generalised in various ways.

Like quantile regression, the representation index chararacterizes the conditional income distribution. However, the representation index can reveal quite different patterns. In the example above, where the conditional top decile cutoff of earnings is $10,000 lower for women than men, it is still possible that the representation of women in the top decile of population earnings could be nearly 10 per cent. This could occur if the highest-performing women earned exactly what men earned, but the next group of women earned much less. The representation index provides direct information on the object of interest: the degree to which a definable group of individuals is represented in a region of the income distribution.

Under-representation can take many forms. Disadvantaged group members might be clustered close to the anchoring quantile or far from 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 an anchoring value.

Representation and severity indices illuminate the conditional distribution of income, rather than just focusing on, for example, the conditional mean. In our empirical work below, we show that although previous research (see Pendakur and Pendakur (2007)) has shown that Aboriginal men face conditional mean earnings disparity of nearly 50 per cent, the representation index shows that 6.7 per cent of Aboriginal men are in the top decile of the conditional population earnings distribution. These two numbers give very different senses of Aboriginal earnings disparity. Aboriginal men have extremely low average earnings, but only somewhat poor access to the top of the earnings distribution. Thus, policy designed to raise Aboriginal representation in 'good jobs' may have only limited effect on their average earnings.

Our indices may be of direct policy interest. For example, the Canadian Employment Equity Act of 1986 (Employment and Immigration Canada (1989)) is designed to guarantee equal access to employment opportunities in federal government jurisdictions (Pendakur (2000)). Subsequent revisions and court cases have established that equal access should apply throughout the job

classification hierarchy. To the extent that the job hierarchy approaches a continuum of ordered types, or that earnings indicate position in the hierarchy, our representation indices can provide a direct measure of compliance with employment equity legislation. In contrast, wage gaps at various quantiles are only weakly informative of access to employment opportunities.

2 Representation

Let $i = 1, ..., N$ index individuals in a sample. Each individual is member of a group $j = 1, ..., J$ with N_j members. We use y to denote income and X to denote a vector of individual characteristics. These methods are generic in the sense that y could be any ordered discrete or continuous variable of interest, such as wages, socio-economic status or education. Denote the joint density of income and characteristics among members of group j by $f_j(y, X)$, and let $f(y, X)$ be the joint density of an anchoring group. We call $f(y, X)$ the anchoring density. The anchoring group could be a population sub-group (e.g., majority workers), or the population as a whole.

Representation is the proportion of group j's members whose income is below a particular quantile of the anchoring density. Suppose, for example, that women comprise the group of interest. The anchoring group could be men, in which case female representation is the proportion of women in a region of the male income distribution. Alternately, the anchoring group could be the entire population, in which case female representation is the proportion of women in a region of the population distribution of income.

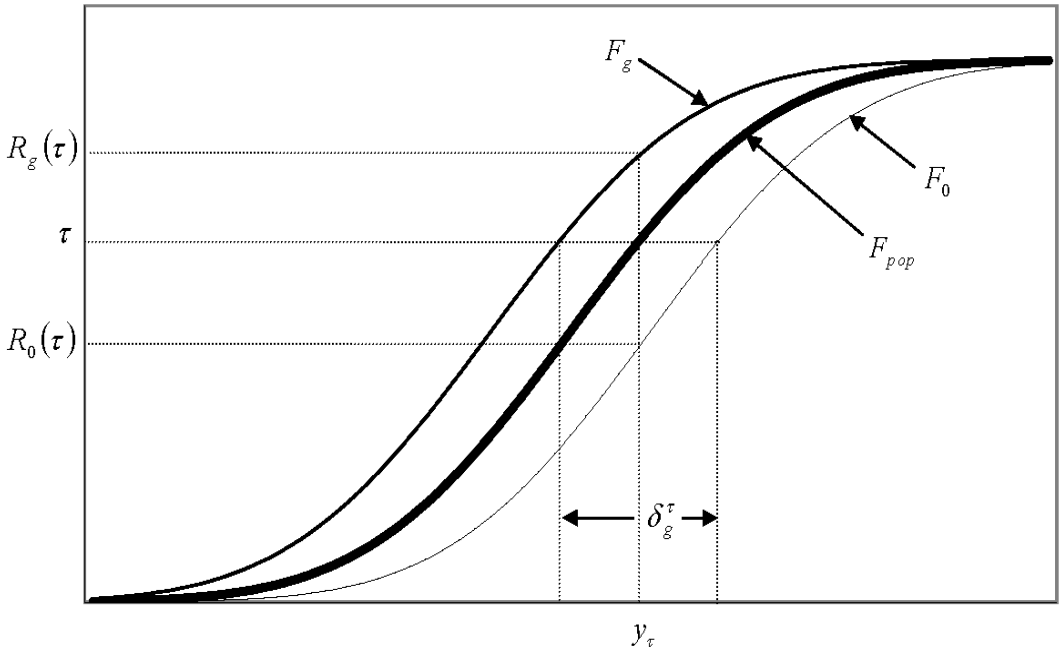
To fix ideas, we begin with an unconditional representation measure. The τ^{th} unconditional quantile of the anchoring distribution, $q(\tau)$, solves $\Pr[y_i < q(\tau)] = \tau$ for i in the anchoring group. Let $\hat{q}(\tau)$ denote a sample estimate of this quantity. In general, $\hat{q}(\tau)$ can be obtained nonparametrically, e.g., by sorting the data on y. Our unconditional representation index for group j, $R_j(\tau)$, is the proportion of group members whose income lies below the τ^{th} unconditional quantile of the anchoring distribution. That is, $R_j(\tau) = F_j(q(\tau))$, where $F_j(y)$ is the cumulative distribution function (cdf) of earnings for group j. A sample estimate is:

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

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

Consider the representation of group j in the bottom decile of the anchoring distribution. By definition, ten percent of the anchoring group earn less than

Figure 1: Quantile Differentials and Representation



$\hat{q}(0.1)$. If $\hat{R}_j(0.1) = 0.2$, then twenty percent of group j's members earn less than $\hat{q}(0.1)$. Members of group j are over-represented by a factor of two in the bottom decile of the anchoring distribution.

Figure 1 illustrates the relationship between quantile wage differentials and the unconditional representation index. The figure shows the cumulative distribution function of income for a hypothetical population (F_{pop}), a reference group (F_0), and a minority group (F_g). At a given quantile τ, the quantile wage differential between group g and the reference group, δ_g^τ, is the horizontal distance between F_g and F_0. At the τ^{th} population quantile, y_τ, the vertical distance between F_g and F_0 is the difference between representation of the two groups. Thus representation and quantile cutoffs convey related, but different, features of the distribution of interest.

Over-representation at the bottom, or under-represenation at the top, could reflect discrimination, or it might arise because group j's members have "bad" characteristics. It will therefore be of interest to compare unconditional representation to a conditional measure, so we can assess the extent to which under- or over-representation is due to individual characteristics.

Let $F_j(y|X)$ denote the conditional cdf of y given X for members of group

j, and let $F(y|X)$ denote the conditional cdf for the anchoring group. The τ^{th} quantile of the anchoring distribution of y conditional on X, $q(\tau, X)$, is the inverse of the conditional cdf: $q(\tau, X) = F^{-1}(\cdot|X)$. Since $F(q(\tau, X)|X) = \tau$ by definition, we have:

$$\frac{\int_0^{q(\tau,X)} f(y, X)dy}{\int_0^{\infty} f(y, X)dy} = \tau. \tag{2}$$

We define the conditional representation function, $r_j(\tau, X)$, as the proportion of group j's members with characteristics X whose income is below the τ^{th} conditional quantile of the anchoring distribution:

$$r_j(\tau, X) = F_j(q(\tau, X)|X). \tag{3}$$

If, for some value of X, this quantity exceeds τ, group members with characteristics X are over-represented in the region below τ, as compared to the anchoring group. If it is less than τ, then the group is under-represented in that region.

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. However, this lack of structure comes at a price. Evaluating $r_j(\tau, X)$ for any value of X is a nonparametric problem that may have a 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 a particular X is not revealing of representation for the group as a whole. This motivates a summary statistic that averages $r_j(\tau, X)$ over group members.

Averaging $r_j(\tau, X)$ over individuals implicitly averages over their characteristics with weights corresponding to the distribution of X in group j. We denote this average as $r_j(\tau)$, and call it the *representation index*:

$$r_j(\tau) = \frac{1}{N_j} \sum_{i \epsilon j} r_j(\tau, X_i). \tag{4}$$

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

Comparing the representation index (4) and the unconditional representation (1) is useful to assess how individual characteristics contribute to a group's

over- or under-representation in a region of the income distribution. Consider, once again, representation in the bottom decile. The value of $\hat{r}_j(0.1)$ gives the proportion of group j's members whose income is in the bottom decile of the anchoring distribution, given their characteristics X. In contrast, the value of $\hat{R}_j(0.1)$ gives the proportion of the group's members whose income is in the bottom decile of the unconditional anchoring distribution. If $\hat{R}_j(0.1) = 0.2$ and $\hat{r}_j(0.1) = 0.1$, then poor characteristics explain the over-representation of group j in the bottom decile of the unconditional anchoring distribution. On the other hand, if $\hat{R}_j(0.1) = 0.2$ and $\hat{r}_j(0.1) = 0.15$, poor characteristics do not explain all of the group's over-representation in the bottom decile. Controlling for their characteristics, they remain over-represented by 50 percent.

Note that the conditional representation function, $r_j(\tau, X)$, and the anchoring conditional quantile function, $q(\tau, X)$, completely characterize the joint distribution of (y, X) for 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 for each group, $q_j(\tau, X)$.

Typically, quantile functions are not estimated for all possible values of τ; rather, they are estimated for a sparse grid of τ values, or even a single τ. Thus, an important question is whether we learn more about a group's representation in a region of the income distribution from the representation index or quantile function at a *single value* of τ. We argue that the representation index more directly illuminates the object of interest. Consider a simple example. If the representation of minority workers in the top decile of income is 0.06, then there are 40 percent fewer minority workers in the top decile of income than we would expect given their characteristics X. In contrast, if we focus solely on estimated quantiles and find that the top decile of minority workers' income 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 index provides direct information on the object of interest: the degree to which a definable group of individuals is represented in a region of the income distribution.

The representation index is nonparametric. If the dimensionality of X is low, it can be estimated using nonparametric estimates of F and F_j. In cases where the dimensionality of X is too large for a nonparametric approach, minimal parametric structure can facilitate estimation. One possibility is to use nonparametric single (or multiple) index density estimates. Quantile regression is an easily implemented alternative, as we now show.

It is straightforward to estimate the representation index, $r_j(\tau)$, in two steps using popular statistical software. The first step is to estimate conditional quantiles for the anchoring group from the quantile regression of y on

X. The anchoring conditional quantile *regression* function, $Q(\tau, X)$, satisfies $\Pr[y_i < Q(\tau, X)] = \tau$ for i in the anchoring group.[1] Second, use the estimated conditional quantile regression function to construct predicted values $\hat{Q}(\tau, X_i)$ for each i in group j. A sample estimate of the representation index for group j is $\hat{r}_j(\tau) = N_j^{-1} \sum_{i \epsilon j} I\left[y_i < \hat{Q}(\tau, X_i)\right]$ 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] A Stata .ado file that estimates the representation index using quantile regression is available at www.sfu.ca/∼pendakur.

3 Severity of Under-Representation

It is natural to ask whether over- or under-representation in a region of the income distribution has large or small consequences. For example, if minorities are under-represented in the bottom decile but over-represented below the 5th percentile, then the representation index at the 10th percentile, $r_j(0.1)$, will be below 0.1, but minority workers might still be meaningfully crowded into the bottom of the distribution. In this section, we present a *severity index* which aggregates, or summarizes, representation across the quantiles. Thus, if the researcher is interested mainly representation in the bottom decile, s/he could estimate the representation index for the bottom decile, and supplement this with the severity index which aggregates representation at quantiles below that. The severity index should focus on the quantile of interest, and should respond strongly to over-representation at quantiles far below. A natural way to aggregate for this purpose is to weight representation below (above) a cutoff by some function of dollar distances from the cutoff. This idea is similar to Sen's (1976) proposal to weight poverty indices by a function of dollar distances from poverty cutoffs (see also Foster et al. (1984)).

The *expectile* function (see Newey and Powell (1987)) defines a convenient way to weight representation. It can be expressed as the solution to a weighted quantile problem. The quantile function 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 is the dollar value of the distance from

[1] We use Q rather than q to denote the quantile conditional regression function because quantile regression imposes parametric structure on the problem, even though q is a non-parametric object.

[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.

the cutoff. The expectile function, e, is thus defined by:

$$\frac{\int_0^{e(\tau,X)} |e(\tau,X)-y|\ f(y,X)dy}{\int_0^{\infty} |e(\tau,X)-y|\ f(y,X)dy} = \tau, \tag{5}$$

which simply adds the weight $|e(\tau,X)-y|$ to (2). 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 $\tau = 0.5$, the total shortfall equals the total surplus, which characterizes the mean. Thus $e(0.5,X)$ is the conditional mean, which can be estimated by ordinary least squares. Expectiles for other τ can be estimated by weighted least squares.

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

$$s_j(\tau,X) = \frac{\int_0^{e(\tau,X)} |e(\tau,X)-y|\ f_j(y,X)dy}{\int_0^{\infty} |e(\tau,X)-y|\ f_j(y,X)dy}. \tag{6}$$

Note that $s_j(\tau,X) = \tau$ for all τ if and only if $f_j(y,X) = f(y,X)$.

Our severity function has a natural metric. For a given X, the severity function evaluated on the anchoring group equals τ. If $s_j(\tau,X)$ is greater (less) than τ, then the dollar-weighted representation of group j below the τ^{th} anchoring expectile is greater (less) than the anchoring group.

The conditional severity function usefully supplements the conditional representation function. For example, if $r_j(0.1,X) = 0.2$ then the proportion of group j's members, with characteristics X, in the bottom decile of the conditional anchoring distribution is twice that of the anchoring group. However, if the earnings of members of group j are clustered just below the bottom decile cutoff, then this over-representation is not very severe. We might find that $s_j(0.1,X) = 0.15$, indicating that when weighted by dollars, over-representation in the bottom of the distribution is not as severe as the conditional representation function suggests.

In this example, we considered representation and severity with $\tau = 0.1$. In general the dollar value of the τ^{th} quantile will not equal the dollar value of the τ^{th} 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 this way has no natural metric. In particular, its value for group j is only meaningful relative to its value for the anchoring group, which does not generally equal τ. In addition, such a measure of conditional severity is not unique, because $q(\tau, X)$ is not unique and hence neither are distances from $q(\tau, X)$.

Like the representation function, $s_j(\tau, X)$ depends on X. A summary measure of severity that averages over X is desirable, so we define the *severity index*, $s_j(\tau)$, as

$$s_j(\tau) = \frac{1}{N_j} \sum_{i \epsilon j} s_j(\tau, X_i). \tag{7}$$

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

Clearly, the choice of weights matters in the severity index. However, whereas Sen's weights in his application to poverty measurement come directly from a social welfare function, in our application the weights lack a corresponding theoretical basis. However, in an application to the income distribution, the dollar distance seems a natural—though *ad hoc*—weight. One could use different (monotone) functions of dollar distances, such as the square root or the natural logarithm, as weights simply by replacing y with that monotone function of y. In our application below, we use log-dollar distances to maintain the spirit of comparability with the literature on log-earnings disparity.

Replacing with sample estimates in (6) and (7) defines a sample estimate of the conditional severity index, $\hat{s}_j(\tau)$, that is easily estimated in two steps. The first step is to estimate the expectile function using the expectile *regression* function $E(\tau, X)$ of the anchoring group.[3]

Expectile regression is related to both ordinary least squares and quantile regression (see Newey and Powell (1987), especially footnote 2).[4] It is based on

[3] Again, we use the notation E rather e because expectile regression imposes parametric structure, even though the expectile function is a nonparametric object.

[4] The difference between these methods is most easily understood as a difference between the penalty function applied to deviations of y_i from a function, $g(\theta, X_i)$, that depends on parameters θ and covariates X. Defining residuals $u_i = y_i - g(\theta, X_i)$, all three methods 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 $p(u) = u^2$. In quantile regression, the penalty function is $p(u) = |\tau - I(u < 0)| \cdot |u|$. In expectile regression, the penalty function is $p(u) = |\tau - I(u < 0)| \cdot u^2$. See Abdous and Remillard (1995) for conditions where quantiles

iterated asymmetrically weighted least squares. Estimation is as follows: given a pre-estimate of the regression function, 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 is repeated to convergence, and the resulting regression model is the estimated expectile regression function, $\hat{E}(\tau, X)$.

The second step is to construct predicted values $\hat{E}(\tau, X_i)$ for all members of group j. A sample estimate of the severity index is the sample average of weighted representation below $\hat{E}(\tau, X_i)$:

$$\hat{s}_j(\tau) = \frac{\sum_{i \epsilon j} \max\left\{\hat{E}(\tau, X_i) - y_i, 0\right\}}{\sum_{i \epsilon j} |\hat{E}(\tau, X_i) - y_i|}. \tag{8}$$

We provide a Stata .ado file that estimates the severity index by this method at www.sfu.ca/∼pendakur.

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

$$\frac{\sum_{i=1}^{N} \max\{\hat{e}(\tau) - y_i, 0\}}{\sum_{i=1}^{N} |\hat{e}(\tau) - y_i|} = \tau. \tag{9}$$

Again, since $e(\tau)$ does not depend on X, neither expectile regression nor parametric structure is needed to estimate $\hat{e}(\tau)$. We can simply sort the data by Y and solve for $\hat{e}(\tau)$. A sample estimate of the unconditional severity index for group j is

$$\hat{S}_j(\tau) = \frac{\sum_{i \epsilon j} \max\{\hat{e}(\tau) - y_i, 0\}}{\sum_{i \epsilon j} |\hat{e}(\tau) - y_i|}. \tag{10}$$

As in the case of representation, 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.

4 Application

We estimate the representation and severity indices on the universe of long form responses to the 2001 Census of Canada. Census long forms are administered to twenty percent of Canadian households, except on Aboriginal reserves

and expectiles coincide.

where all households are surveyed. All reported estimates use sample weights provided by Statistics Canada.[5] We simulate standard errors using the bootstrap. Simulated standard errors for estimates in Tables 1 and 2 are all less than 0.002. Given the precision of our estimates, we omit standard errors from the Tables to minimize clutter. Details are available on request.

We define three broad ethnic categories of interest: Aboriginal, visible minority and white. These categories correspond to those used in Canadian federal 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.

We focus on the native-born population to eliminate the potentially confounding effects of immigration. 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 Census data are 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 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. Pendakur and Pendakur (2007) report conditional mean earnings disparity using the same data and controlling for these same characteristics. They found that: comparing to white women, visible minority and Aboriginal women face log-earnings disparity of -0.04 and -0.16, respectively; and, comparing to white men, visible minority and Aboriginal men face log-earnings disparity of -0.09 and -0.42, respectively. We will consider how the representation index compares to these numbers in our discussion below.

Although Statistics Canada guidelines do not allow us to report the exact counts of population groups, our analysis sample contains approximately 900,000 observations each for men and women. Because these are confidential

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

data, we present estimates based on the Public Use Microdata File (PUMF) of the 2001 Census of Canada in the Appendix to permit replication. 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 the PUMF 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.

Our investigation considers men and women separately, and uses the same-gender native-born population as the anchoring group. Table 1 presents the conditional and unconditional representation index at the tenth, fiftieth, and ninetieth percentile of log earnings. At each quantile, the representation of white men and women corresponds very closely to that of the entire male or female (native-born) population. This is unsurprising, given that white men and women comprise over 95 percent of the native-born. 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.

We begin a closer inspection of Table 1 with the least extreme group, female visible minorities. Compared to the population of women, female visible minorities are unconditionally over-represented by almost 50 percent in the bottom decile of log earnings ($\hat{R}_j(0.1) = 0.149$), and under-represented by nearly 20 percent in the top decile ($\hat{R}_j(0.9) = 0.919$). However, these values are almost completely explained by the characteristics of group members ($\hat{r}_j(0.1) = 0.104, \hat{r}_j(0.9) = 0.904$). This is similar in spirit to the very small conditional mean log-earnings disparity reported in Pendakur and Pendakur (2007).

Male 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 ($\hat{R}_j(0.1) = 0.226$), and 41 percent fewer above the ninetieth percentile, than in the population ($\hat{R}_j(0.9) = 0.941$). 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. Thus, some under-representation remains in the top decile of earnings: only three-quarters as many visible minorities in are this region as would be expected if representation were 'fair'.

Table 1: Representation Index for Selected Demographic Groups

	Unconditional			Conditional		
	$\tau = .1$	$\tau = .5$	$\tau = .9$	$\tau = .1$	$\tau = .5$	$\tau = .9$
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.

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 worse 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: the representation index shows that Aboriginal women remain over-represented in the bottom decile by 42 percent, and under-represented in the top decile by 18 percent. In contrast, for Aboriginal men, controlling for characteristics does not much change the over-representation at the bottom, but reduces the under-representation at the top by half: Aboriginal men are under-representated by 33 per cent ($\widehat{R}_j(0.9) = 0.933$) in the top decile.

The results on representation for Aboriginal men are striking. They face mean log-earnings disparity of -0.42, which is more than four times the disparity of -0.09 faced by visible minorities (and much larger than that faced by Black men in the USA). But, the representation index shows that 6.7 per cent of Aboriginal men are found in the top decile of the population conditional earnings distribution, which is not so different from the 7.6 per cent

Table 2: Severity Index for Selected Demographic Groups

	Unconditional			Conditional		
	$\tau = .1$	$\tau = .5$	$\tau = .9$	$\tau = .1$	$\tau = .5$	$\tau = .9$
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.

of visible minorities found in this region. Thus, the 'traditional' conditional mean measure and the representation index illuminate very different aspects of the conditional earnings distribution. In particular, although Aboriginal men have extremely low conditional mean earnings, they have only somewhat poor access to the top of the earnings distribution. An implication of this is that policy which targets the access of Aboriginal men to 'good jobs' may have only a small effect on their mean earnings.

Table 2 presents estimates of the severity index. For most groups, they are qualitatively similar to the representation index. However, we see that the representation index substantially understates the poor outcomes of Aboriginal men. For this group, the unconditional severity index at the mean, $\hat{S}_j(0.5)$, is 0.856. This is more dismal than their (unweighted) representation below the median of log earnings, $\widehat{R}_j(0.5) = 0.705$, because the earnings of Aboriginal men are concentrated in the lowest part of the log earnings distribution. Accounting for the characteristics of Aboriginal males mitigates the severity of over-representation somewhat: $\widehat{s}_j(0.5) = 0.731$. Indeed, even at the bottom decile, the severity index is larger than the representation index, suggesting that the earnings of Aboriginal men are more crowded into the lower part of that region of the distribution. Thus, the severity index usefully supplements

the representation index: for Aboriginal men, the severity index suggests that over-representation in the bottom half or decile is exacerbated by crowding at the bottom.

5 Conclusion

The representation index provides an intuitive and easily computed measure of a group's representation in a region of the income 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. 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 wage outcomes of minority groups.

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. The evidence suggests that these groups face some exclusion from society's best jobs, and are crowded into employment in society's worst jobs.

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	
Grades 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	6.38		8.38	
University, with bachelor or first professional degree	12.9		15.3	
University, with university certificate above bachelor	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, Public Use Microdata File

	Unconditional			Conditional		
	$\tau = .1$	$\tau = .5$	$\tau = .9$	$\tau = 0.$	$\tau = .5$	$\tau = .9$
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 of the 2001 Census of Canada. Simulated standard errors are available on request. All standard errors are less than 0.006.

Appendix: Replicability

To allow replication, we estimate the representation and severity indices on the Public Use Microdata File (PUMF) of the 2001 Census of Canada. Estimates 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). There are some 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. 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.

Appendix Table 3: Severity Index for Selected Demographic Groups, Public Use Microdata File

	Unconditional			Conditional		
	$\tau = .1$	$\tau = .5$	$\tau = .9$	$\tau = .1$	$\tau = .5$	$\tau = .9$
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 of the 2001 Census of Canada. Simulated standard errors are available on request. All standard errors are less than 0.003.

References

Abdous, A. and B. Remillard (1995). Relating quantiles and expectiles under weighted-symmetry. *Annals of the Institute of Statistical Mathematics 47*(2), 371–384.

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.

Employment and Immigration Canada (1989). *Employment Equity Act Annual Report*. Ottawa, ON, Canada: Minister of Supply and Service.

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

Foster, J., J. Greer, and E. Thorbecke (1984). A class of decomposable poverty indices. *Econometrica 52*, 761–766.

Kopczuk, W., E. Saez, and J. Song (2007). Uncovering the American dream: Inequality and mobility in Social Security earnings data since 1937. NBER Working Paper No. 13345.

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 (2007). Minority earnings disparity across the distribution. *Canadian Public Policy 33*(1), 41–62.

Pendakur, R. (2000). *Immigrants in the Labour Force: Policy, Regulation and Impact*. Montreal, PQ, Canada: McGill-Queens University Press.

Sen, A. (1976). Poverty: An ordinal approach to measurement. *Econometrica 44*(2), 219–231.

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.