Oxford Poverty & Human Development Initiative (OPHI)
Oxford Department of International Development
Queen Elizabeth House (QEH), University of Oxford

OPHI
Research in Progress series 2010

This paper is part of the Oxford Poverty and Human Development Initiative's Research in Progress (RP) series. These are preliminary documents posted online to stimulate discussion and critical comment. The series number and letter identify each version (i.e. paper RP1a after revision will be posted as RP1b) for citation.

Please cite these papers as Author Last name, First name, "Title" (Year) *OPHI Research in Progress ##a.* (For example: Alkire, Sabina and Foster, James "Counting and Multidimensional Poverty" (2007) *OPHI Research in Progress 1a.*)

For more information, see **www.ophi.org.uk**.

A comparison between the Pearson-based dissimilarity index and the multiple-group overlap index[*]

Gaston Yalonetzky[†]

Abstract

The Overlap Index has received renewed attention as a measure of inequality between two distributions (e.g. Anderson, Ge, and Leo, 2010). In this paper I propose a straightforward extension of the Overlap index to comparisons of multinomial, multivariate distributions of wellbeing over several groups. I also compare its properties against those of a dissimilarity index based on Pearson's goodness of fit statistic. The comparison highlights that both indices are similar in declaring perfect between-group equality if and only if the distributions are identical. They are diferent in their sensitivity to most migrations of individuals from one wellbeing state to another (with meaningful exceptions), and to different population invariance axioms. They also differ in the situations under which they attain its value of maximum inequality. The comparison illustrates the importance of both group size and concepts of minimum and maximum inequality in between-group inequality analysis. An empirical application of both indices, to a cohort history of between-group inequality in educational attainment in India, shows that such inequality across gender and caste increased toward independence and then, after the 1950s, it started to decline.

JEL Codes: C10
Keywords: Between-group inequality

1 Introduction

A significant part of the literature measuring between-group inequality relies on indices that either compare standards of the group distributions (e.g. the means) or compare the sample values directly.[1] Another part of the literature measures between-group inequality

---

[1]A good example of measures comparing standards of the group distributions is the between-group inequality component yielded by decompositions of path-independent indices into between and within-group components. A complete discussion of these measures is in Foster and Shneyerov (2000). A recent example of a refined between-group inequality measure based on comparisons of distributional standards is provided by Elbers, Lanjouw, Mistiaen, and Ozler (2008). Examples of between-group indices comparing the sample values abound. For instance Ebert (1984); Dagum (1987); and several examples from the polarization literature (e.g. see Esteban and Ray, 1994; Duclos, Esteban, and Ray, 2004).

on probability space.[2] Examples of the latter are the overlap index (e.g. Weitzman, 1970; Anderson, Ge, and Leo, 2010), the dissimilarity index of the Human Opportunity Index (HOI) (Barros, Ferreira, Molinas, and Saavedra, 2009), the Pearson-based dissimilarity index (Yalonetzky, 2009) and indices based on relative distributions (e.g. Handcock and Morris, 1999; Breton, Michelangeli, and Peluso, 2008; Yalonetzky, 2010).

In this paper I discuss the similarities and differences between two such indices: a multi-group, multivariate version of the overlap index and the Pearson-based dissimilarity index. Why is this comparison relevant? Firstly, discrete-variable versions of both indices are suitable to analise between-group inequality over multinomial and multivariate distributions of wellbeing. The multinomial measurement of several indicators of wellbeing enhances the practical relevance of such indices.[3] Secondly, in certain contexts, certain values of both indices reflect relevant concepts. For instance, in the analysis of inequality of opportunities both indices achieve their minimum value if and only if distributions of wellbeing are identical, which reflects, for instance, so-called *circumstance neutralization* (Fleurbaey, 2008). [4] Few other indices in the inequality of opportunity literature share this property.[5]

When comparing both indices, I find that besides the aforementioned similarity, they react in the same way when a migration of one individual within one group between two (discrete) states of wellbeing restores (or breaks) the equality of the probability of attaining those two states across groups. I also find that the indices are different in three crucial aspects: General sensitivity to intra-group migrations between states; fulfillment of different population invariance axioms; and situations in which they declare maximum between-group inequality. These differences are worth higlighting also because they inform the researcher interested in performing between-group inequality analysis on the choice of indicator, in the context of multinomially distributed variables. For instance, if the researcher wants the index to be sensitive to variation in group sizes then the Pearson-based dissimilarity index is preferrable. Otherwise the overlap index, being insensitive to heterogeneous variations in group size, should be used.

This paper uses the two indices in an interesting empirical application to document a cohort history of between-group inequality in educational attainment in India. According to both indices, educational inequality across gender and caste increased toward independence and soon thereafter. But then, after the 1950s, it started to decline. According to the overlap index this decline has been monotonic, whereas the Pearson-based index shows an increase from the second-to-youngest to the youngest cohort. Yet the latter increase does not counter the significant reduction in between-group inequality revealed by the same index.

The paper proceeds as follows: The next section introduces the basic notation and definitions, followed by a section presenting the indices and their similarities. Then follows

---

[2]The Gini coefficient is one special case in which estimation of it can be made either on the space of the variable's values or on that of probabilities.

[3]For a good example of the multinomial measurement of several indicators of different aspects of well-being see the questionnaire of the recent Chile Missing Dimensions Dataset collected by the Oxford Poverty and Human Development Institute. (OPHI and de Chile, 2009)

[4]Fleurbaey (2008) defines circumstance neutralization as an allocation of resources in which " [...] it should be possible to express individual well-being as a function of responsibility characteristics only." (Fleurbaey, 2008, p. 26)

[5]One notable exception is the within-tranch approach of Checci and Peragine (2005).

a detailed discussion of the indices' main differences, followed by the empirical application. The paper finally concludes with some remarks.

2 Basic notation, definitions and axioms

A contingency table, M, is defined as a matrix with A rows and T columns, where $(A, T) \in \mathbb{N}^2_{++}$, and whose elements belong to the set of non-negative, natural numbers, representing absolute frequencies or observations. A typical element of a contingency table in row α and column t, denoting the number of observations in those coordinates, is: N^t_k. The sum of all elements of column t is: $N^t \equiv \sum_{\alpha=1}^{A} N^t_\alpha$. Likewise the sum of all elements in row α is: $N_\alpha \equiv \sum_{t=1}^{T} N^t_\alpha$. And the total number of observations distributed across the table is: $N \equiv \sum_{\alpha=1}^{A} \sum_{t=1}^{T} N^t_\alpha$.

The following definitions are necessary for the next sections of the paper:

- The percentage of total observations in column t: $w^t \equiv \frac{N^t}{N}$.

- The probability of being in row α conditional on belonging to column t: $p^t_\alpha \equiv \frac{N^t_\alpha}{N^t}$.

- The probability of being in row α: $p^*_\alpha \equiv \frac{N_\alpha}{N} = \sum_{t=1}^{T} w^t p^t_\alpha$.

Definition 1 *A minimum intra-column migration (MIM): A contingency table $\widehat{M}$ is obtained from a table M by MIM if there are rows i and j, and column τ, such that: $\widehat{p^t_\alpha} = p^t_\alpha \forall t \wedge \forall \alpha \neq (i \vee j)$; $\left(\widehat{p^t_i} = p^t_i \wedge \widehat{p^t_j} = p^t_j \right) \forall t \neq \tau$; $\widehat{p^t_i} = p^t_i - \delta \wedge \widehat{p^t_j} = p^t_j + \delta$; where $0 < \delta \leq p^t_i$.*

In the analysis of between-group inequality, the definition (MIM) is helpful in analyzing intra-groups migrations of individuals from one wellbeing state to another. In such a context, the table's columns stand for different groups (e.g. T types in inequality of opportunity analysis) and the rows stand for discrete, multinomial wellbeing outcomes (e.g. A states). Using this interpretation of rows and columns of a table, $\mathcal{M}_{TA}$ is defined as the set of all tables with a fixed number of groups T and a fixed number of states A. The set of all contingency tables with a fixed number of groups, but for any number of states is denoted by $\mathcal{M}_{T*} = \cup_{\mathbf{A} \subset \mathbb{N}} \mathcal{M}_{TA}$. Similarly, the set of all tables with a fixed number of state, but for any number of groups is denoted by $\mathcal{M}_{*A} = \cup_{\mathbf{T} \subset \mathbb{N}} \mathcal{M}_{TA}$. Finally, the set of all tables is denoted by $\mathcal{M} = \cup_{\mathbf{T}, \mathbf{A} \subset \mathbb{N}} \mathcal{M}_{TA}$. A between-group inequality indicator, $I(\mathcal{M})$, is a mapping from a contingency table to a real number: $I(\mathcal{M}) : \mathcal{M} \to \mathbb{R}$. In this paper I focus on a class of between-group inequality indicators normalized to have values between zero and one: $C(\mathcal{M}) \subset I(\mathcal{M}) \mid C(\mathcal{M}) : \mathcal{M} \to [0, 1]$.

An important concern regarding these indicators is whether their value is affected by the population size and by the relative sizes of the groups. In terms of the contingency table the question is whether the indicators are sensitive to N, on one hand; and to differential changes in w^t on the other hand. The following two axioms are related to these sensitivities:

Axiom 1 *Population invariance (PI): if a table $\widehat{M}$ is obtained from a table M by a replication of the whole population, such that $\widehat{N^t_\alpha} = \lambda N^t_\alpha \forall t \in [1, T], \alpha \in [1, A] \wedge \lambda > 0$; and $C\left(\widehat{M}\right) = C(M)$, then C is said to be population invariant.*

Axiom 2 *Group composition invariance (GI) if a table $\widehat{M}$ is obtained from a table M by differential replications of the population's groups, such that $\widehat{N_\alpha^t} = \lambda^t N_\alpha^t \forall t \in [1, T]$, $\alpha \in [1, A] \wedge \lambda^t > 0$; and $C\left(\widehat{M}\right) = C(M)$, then C is said to be group composition invariant.*

With these definitions and axioms now the indices are introduced and compared to each other.

3 The indices and their similarities

The Pearson-based dissimilarity index is based on a statistic proposed by Anderson and Goodman (1957) to test the null hypothesis of homogeneity of multinomial distributions. Its formula is:

$$H = \frac{1}{\min\{T-1, A-1\}} \sum_{t=1}^{T} \sum_{\alpha=1}^{A} w^t \frac{(p_\alpha^t - p_\alpha^*)^2}{p_\alpha^*} \tag{1}$$

The multivariate, multiple-group overlap index is based on the two-group overlap index proposed by Weitzman (1970). Its formula is:

$$O = 1 - \sum_{\alpha=1}^{A} \min\{p_\alpha^1, p_\alpha^2, ..., p_\alpha^T\} \tag{2}$$

Both indices have the key similarity of attaining a minimum value of zero if and only if the group distributions are all identical. That is: $(H = O = 0) \leftrightarrow (p_\alpha^i = p_\alpha^j = p_\alpha^* \forall i \neq j)$. The proof of this proposition is straightforward.

Another similiarity between the two indices is that they both strictly decrease whenever a MIM restore pairwise state between-group equality (PSBE) and strictly increase when they break PSBE. The latter is defined by: $\exists (\alpha, \beta) \in [1, ..., A] \mid p_\alpha^1 = p_\alpha^2 ... = p_\alpha^T = p_\alpha^* \wedge p_\beta^1 = p_\beta^2 ... = p_\beta^T = p_\beta^*$. A MIM that restores PSBE is characterized by: $\widehat{p_i^1} = \widehat{p_i^2} = ... = \widehat{p_i^T} = \widehat{p_i^*} \wedge \widehat{p_j^1} = \widehat{p_j^2} = ... = \widehat{p_j^T} = \widehat{p_j^*}$, i.e. both the dispatching and the receiving state are rendered homogeneous across groups after the MIM. By contrast, a MIM that breaks PSBE is characterized by: $p_i^1 = p_i^2 = ... = p_i^T = p_i^* \wedge p_j^1 = p_j^2 = ... = p_j^T = p_j^*$. The proof that H strictly decreases whenever a MIM restores PSBE and strictly increases whenever a MIM breaks PSBE is in Yalonetzky (2009). The proof for O requires noticing first that a MIM impacts on the index the following way:

$$\Delta O \equiv O\left(\widehat{M}\right) - O(M) = -\min\left\{p_i^1, ..., p_i^\tau - \delta, ..., p_i^T\right\} - \min\left\{p_j^1, ..., p_j^\tau + \delta, ..., p_j^T\right\} \tag{3}$$
$$+ \min\left\{p_i^1, ..., p_i^\tau, ..., p_i^T\right\} + \min\left\{p_j^1, ..., p_j^\tau, ..., p_j^T\right\}$$

Equation (3) accounts already for the fact that in a MIM the migration does not affect other groups' probabilities. If a MIM restores PSBE then: $\min\left\{p_i^1, ..., p_i^\tau - \delta, ..., p_i^T\right\} = \min\left\{p_i^1, ..., p_i^\tau, ..., p_i^T\right\}$ (because $\widehat{p_i^\tau} = p_i^*$) and $\min\left\{p_j^1, ..., p_j^\tau + \delta, ..., p_j^T\right\} > \min\left\{p_j^1, ..., p_j^\tau, ..., p_j^T\right\}$ (because

$\widehat{p_j^\tau} = p_j^*$). Hence a MIM that restores PSBE implies $\Delta O < 0$. Reversing the migration proves that a MIM that breaks PSBE also strictly increases the value of the overlap index.

Therefore the indices agree in declaring perfect between-group inequality if and only if distributions are homogeneous across groups and unanimously react when pairwise equality of states is restored or broken by MIM. For other effects of MIM, however the indices may disagree in their sensitivity. That fact along with fulfillment of different population invariant axioms and attainments of maximum inequality are the three main differences that are discussed in the next section.

4   The differences

4.1   Sensitivity to intra-group migrations

Even though both indices react similarly to MIMs that restore (or break) PSBE, they have differential sensitivity to MIMs that neither restore or break PSBE. In principle, the latter type of migration ought to have an a priori ambiguous effect on either index because such migration may increase or decrease the heterogeneity across the group-specific probabilities of being in the departing state as well as increase or decrease the heterogeneity across the group-specific probabilities of being in the receiving state.

In the case of O a MIM's effect on the index is described by (3). For the effect to be null it is sufficient (but not necessary) that:

$$\exists m, n \in [1, T] \mid p_i^m \le p_i^\tau - \delta < p_i^\tau \land p_j^n \le p_j^\tau < p_j^\tau + \delta. \tag{4}$$

Alternatively, if condition (4) is not met, the effect can still be null if and only if:

$$p_j^\tau + \delta - \min\left\{p_j^1, ..., p_j^\tau + \delta, ..., p_j^T\right\} = p_i^\tau - \min\left\{p_i^1, ..., p_i^\tau, ..., p_i^T\right\}. \tag{5}$$

If the left-hand side of (5) is higher than the right-hand side then $\Delta O > 0$, otherwise $\Delta O < 0$. Notice that this sensitivity is independent of w^t.

By contrast neither condition (4) nor (5) are necessary or sufficient to render the effect of a MIM on H null. In other words, H is sensitive to several different MIMs fulfilling either (4) or (5). This can be shown with counter-examples but eye inspection of the effect of MIM on H should do:

$$
\begin{aligned}
\Delta H \ \equiv \ H\left(\widehat{M}\right) - H\left(M\right) = &\frac{\delta w^\tau}{\min\{T-1, A-1\}}\{[\frac{\sum_{t=1}^T w^t \left(p_i^t - p_i^*\right)^2}{p_i^* \left(p_i^* - \delta w^\tau\right)} \\
&- \frac{\sum_{t=1}^T w^t \left(p_j^t - p_j^*\right)^2}{p_j^* \left(p_j^* + \delta w^\tau\right)}] \\
&+ 2[\frac{p_j^\tau - p_j^*}{p_j^* + \delta w^\tau} - \frac{p_i^\tau - p_i^*}{p_i^* - \delta w^\tau}] + \delta\left(1 - w^\tau\right)\frac{p_i^* + p_j^*}{\left(p_i^* - \delta w^\tau\right)\left(p_j^* + \delta w^\tau\right)}\}
\end{aligned}
\tag{6}
$$

Condition (6) shows some interesting features. Firstly, that $\Delta H \ne 0$ is possible with several combinations of probabilities and MIM's δ, that fulfill conditions (4) or (5), hence

$\Delta O = 0$. In other words, O is either plainly insensitive to those same combinations (fulfillment of (4)) or their effects cancel out (fulfillment of (5). Secondly it is also possible to find combinations of probabilities and MIM's δ such that $\Delta H = 0$ and $\Delta O \neq 0$.[6] However, by comparing (6) to (4) and (5), it is fair to conclude that it is harder to find combinatios of probabilities and MIM's δ in which $\Delta H = 0$ and $\Delta O \neq 0$ than to find combinations in which $\Delta H \neq 0$ and $\Delta O = 0$. One reason behind this is that ΔH depends on evaluations of many more comparisons between the probabilities than ΔO. In general, considering that a table of group probabilities has $T(A-1)$ degrees of freedom, for a fixed number of groups, T, and states, A, it should always be easier to find cases in which $\Delta O = 0$ than cases in which $\Delta H = 0$. Thereby it should be easier to find cases of $\Delta H \neq 0$ and $\Delta O = 0$ than the other way around. In other words, for a given number of degrees of freedom, H is more sensitive to MIM's than O.

4.2 Fulfillment of population invariance axioms

If a table $\widehat{M}$ is obtained from a table M by a replication of the whole population, such that $\widehat{N_\alpha^t} = \lambda N_\alpha^t \forall t \in [1,T]$, $\alpha \in [1,A] \wedge \lambda > 0$, then $\widehat{p_\alpha^t} = \frac{\widehat{N_\alpha^t}}{\widehat{N^t}} = \frac{\widehat{N_\alpha^t}}{\sum_{\alpha=1}^A \widehat{N_\alpha^t}} = \frac{\lambda N_\alpha^t}{\lambda \sum_{\alpha=1}^A N_\alpha^t} = p_\alpha^t, \forall (\alpha,t) \in [1,A] \times [1,T]$; and $\widehat{w^t} \equiv \frac{\widehat{N^t}}{\widehat{N}} = \frac{\lambda \sum_{\alpha=1}^A N_\alpha^t}{\lambda \sum_{t=1}^T \sum_{\alpha=1}^A N_\alpha^t} = w^t$. Therefore $H(\widehat{M}) = H(M)$ and $O(\widehat{M}) = O(M)$; i.e. both indices fulfill PI.

However if if a table $\widehat{M}$ is obtained from a table M by differential replications of the population's groups, such that $\widehat{N_\alpha^t} = \lambda^t N_\alpha^t \forall t \in [1,T]$, $\alpha \in [1,A] \wedge \lambda^t > 0$, then it is still the case that: $\widehat{p_\alpha^t} = \frac{\widehat{N_\alpha^t}}{\widehat{N^t}} = \frac{\widehat{N_\alpha^t}}{\sum_{\alpha=1}^A \widehat{N_\alpha^t}} = \frac{\lambda^t N_\alpha^t}{\lambda^t \sum_{\alpha=1}^A N_\alpha^t} = p_\alpha^t, \forall (\alpha,t) \in [1,A] \times [1,T]$. Yet now $\widehat{w^t} \equiv \frac{\widehat{N^t}}{\widehat{N}} = \frac{\lambda^t \sum_{\alpha=1}^A N_\alpha^t}{\sum_{t=1}^T \lambda^t \sum_{\alpha=1}^A N_\alpha^t} \neq w^t$ (unless $\lambda^i = \lambda^j = \lambda \forall (i,j) \in [1,T]$). Therefore $O(\widehat{M}) = O(M)$, i.e. the overlap index fulfills GI because it is insensitive to group weights. By contrast, there is no guarantee that the value of the Pearson-based index remains unaffected since it is sensitive to weights directly and indirectly (through the composition of the average probabilities). Therefore this index does not fulfill GI. For instance suppose that $\lambda^t = 1 \forall t \neq \tau \wedge \lambda^\tau = \lambda$. In that case then:

$$\widehat{w^\tau} = \frac{\lambda w^\tau}{\lambda w^\tau + (1 - w^\tau)} \tag{7}$$

$$\widehat{w^t} = \frac{w^t}{\lambda w^\tau + (1 - w^\tau)} \forall t \neq \tau \tag{8}$$

$$\widehat{p_\alpha^*} = \frac{p_\alpha^* + (\lambda - 1) w^\tau p_\alpha^\tau}{\lambda w^\tau + (1 - w^\tau)} \tag{9}$$

With equations (7) through (9) it is possible to establish that: $\lim_{\lambda \to \infty} \widehat{w^\tau} = 1$, $\lim_{\lambda \to \infty} \widehat{w^t} = 0 \forall t \neq \tau$ and $\lim_{\lambda \to \infty} \widehat{p_\alpha^*} = p_\alpha^\tau$. Therefore: $\lim_{\lambda \to \infty} H = 0$; that is to say, $H(\widehat{M}) \neq H(M)$

---

[6]Notice also that the last element to the right-hand side of (6) remains positive even when before the MIM the departing and receiving states, i and j, are homogeneous. That is precisely the proof that a MIM that breaks PSBE increases the value of H. See Yalonetzky (2009).

and H does not fulfill GI.

4.3   Concepts of maximum between-group inequality

Both the Pearson-based dissimiliarity index and the overlap index are bounded between zero and one. The first value is attained if and only if distributions are identical across groups. The second value, which signals the maximum inequality that the indices attribute, is attained by each index under similar but not identical circumstances. Both indices measure inequality as association between the groups (e.g. the columns from 1 to T) and the states of wellbeing (e.g. the rows from 1 to A). Is their value of maximum inequality related to a notion of maximum association? Kendall and Stuart define two notions of extreme high association. The first one, called *complete association*, occurs if all individuals having an attribute A also have an attribute B, even though not everyone having attribute B may have attribute A.[7] The second one, called *absolute association*, occurs when all individuals having attribute A also have attribute B and all those having attribute B also have attribute A (Kendall and Stuart, 1973, chapter 33). These definiton are given in the context of contingency tables with two variables hiavng two states each. More general definitions for two variables and several states/values in each variable, can be proposed. Complete association can be ascertained when all individuals having a value A of row variable have value B of column variable although not all those having value B may have value A.[8] Absolute association is likewise said to occur when all individuals having value A of row variable have value B of column variable and all those having the latter column value have also the former row value.

The overlap index is equal to one if and only if: $\forall \alpha \colon \exists i^\alpha, j^\alpha \in [1, T] \mid p_\alpha^{i^\alpha} = 0 \wedge p_\alpha^{j^\alpha} > 0$. Therefore the overlap index attains its maximum value, for instance, when every state of the multinomial distribution is associated with a subset of groups with strictly fewer elements than the set of all groups. Different situations of complete association can fulfill this criterion as well as situations of absolute assocation (e.g. equal numbers of states and groups and each group exclusively associated with one state only). In fact all situations of absolute association lead to $O = 1$. Hence the overlap index can be used to test the null hypothesis of absolute association. By contrast it can not be used as a test of complete association because certain forms of it can come along with substantial overlap across the distributions. Alternatively, the condition under which the overlap index is maximal can be proposed as a condition itself of maximum association. It can be called minimum overlap maximum association (MOMA) and be defined, in reverse of complete association, as a situation in which all individuals having attribute A do not have attribute B and all those having attribute B do not have attribute A.

The dissimilarity index is equal to one if and only if:

- $T < A$ and there is complete association (where the row variable of the definition has range $[1, A]$) ; or

---

[7]Of course, attributes A and B are a priori, or in theory, non-exclusive (e.g. A may mean being a woman and B may mean being healthy).

[8]The roles of row and column variable may be reversed in a meaningful way as is discussed below when describing the circumstances under which $H = 1$.

- $T > A$ and there is complete association (where now the definition must be read reverting the roles of row and column variables, the latter having a range $[1, T]$) ; or

- $T = A$ and there is absolute association.[9]

Hence the dissimilarit index can be used to test the null hypothesis of complete or absolute association depending on the relationship between T and A.[10]

Now, because $\sum_{\alpha=1}^{A} p_\alpha^t = 1 \forall t$, when $T < A$ with complete association (as defined in the bullet points), then MOMA holds. But the reverse is not true. Therefore when $T < A$: $H = 1 \rightarrow O = 1$. When $T > A$ with complete association (as defined in the bullet points), then MOMA holds.[11] But the reverse is true. Therefore, again, when $T > A$: $H = 1 \rightarrow O = 1$. Finally, when $T = A$ with absolute association, MOMA holds but the reverse is not true. Thereby it is always the case that $H = 1 \rightarrow O = 1$.

5   Empirical application

The use of the two indices, H and O, is now illustrated with an application to between-group inequality of educational achievement in India. The pervasiveness of inequality among different groups of people in India, and the different intensity of that inequality across Indian regions, has long drawn the attention of historians and social scientists. In particular they have sought to understand the roots of, and find solutions to, widespread inequalities based on gender, caste and religion. [12] Following an inequality-of-opportunity approach, I classify people into groups defined by circumstances beyond their control.[13] Thereby I combine two categories of gender and four categories of caste[14] in order to construct eight groups of Indian citizens. [15] Then I focus on their educational attainment and ask whether there have been changes in inequality of opportunity across different cohorts of adults in society. Further details about the data are in the next subsection, followed by the results. .

---

[9] For a discussion of the maximum value of H see Yalonetzky (2009).

[10] Since the dissimilarity index is based on Pearson's goodness-of-fit statistic, its asymptotic standard errors can be derived following Kendall and Stuart (1973, chapter 33). As for the overlap index, its asymptotic standard errors can be derived by extending the two-group case studied by Anderson, Linton, and Whang (2009) to multiple groups.

[11] For this to be true an additional restriction needs to be imposed, namely that $\forall \alpha$: $\max\{p_\alpha^1, p_\alpha^2, ..., p_\alpha^T\} > 0$.

[12] For an historical account in the post-independence period see Chandra, Mukherjee, and Mukherjee (2008). A classical treatment of these inequalities in the Economics literature is in Dreze and Sen (1995). Recently Deshpande (2007, and the author's own work referenced therein) has quantitatively documented gender and caste inequalities over different wellbeing dimensions.

[13] For which, therefore, they can not be held accountable and may be even considered for compensation. See e.g. Roemer (1998), Fleurbaey (2008).

[14] The caste categories are: Scheduled tribe; scheduled caste; other backward castes; other castes.

[15] I could have considered religion as well which usually is "acquired" in the household and, so, originally beyond the individual's control during childhood. However religious conversions are possible in adulthood. Hence I leave it outside to focus on to other categories that are harder to change through autonomous decisions: gender and caste.

5.1 Data

The data come from the Indian 2004 NSS. For the cohort analysis the following eight cohorts, defined by age in 2004, are considered: 30 to 34 years old, 35 to 39, 40 to 44, 45 to 49, 50 to 54, 55 to 59, 60 to 64, and 65 years old or older. By focusing on adults 30 years old or older, the probability of having censored observations, i.e. individuals whose observed educational attainment is different from his/her unobserved final educational attainment (secured after 2004), becomes negligible. The educational attainment variable is multinomial, composed of the following seven ordinal categories: not literate (=1); literate without formal schooling (=2); literate with formal schooling but incomplete primary education (=3); just complete primary education (=4); just complete middle education (=5); just complete secondary education (=6); more than complete secondary education (=7).

The distributions of educational attainment by group and cohort are in tables Table 2 through Table 9 in the Appendix. Sample sizes are on Table 10. The main information drawn from the tables is that numerous aspects of educational attainment have improved for all groups among the younger cohorts vis-a-vis the old ones. For instance, the percentages of illiterate people have steadily declined for all groups while the percentage of people with higher education has increased. First-order stochastic dominance of younger people's distributions over older's is easily spotted, especially when comparing non-adjacent cohorts. Between-group inequality manifests in distributions that are more favourable (e.g. in terms of first-order dominance and/or percentages of illiterate people, people with higher education, average attainment, etc.) to men over women; and more favourable to people belonging to the category of "other castes" vis-a-vis people belonging to scheduled tribe, scheduled caste or "other backward castes". Combined, it appears that across all cohorts, women belonging to the latter three caste categories are the relatively most disadvantaged in terms of educational opportunities (as measured by probabilities of attaining different educational levels).

Has the cross-cohort increase in educational attainment been accompanied by higher or lower between-group inequality? Even though it is true that the nature of the variable forces equality if and when all individuals attain the minimum and/or the maximum levels, most shifts in the distributions are expected to yield a priori ambiguous changes in between-group inequality. Whether such inequality increases or not depends, in part, on the different timings that it takes to different groups of society to benefit from phenomena like the expansion of public (and private) education. In the next sub-section an answer to the question is provided.

5.2 Results

The results appear summarized in table Table 1,[16] and illustrated separately for H and O in figures Figure 1 and Figure 2, respectively. Both indices tell a remarkably similar story: From the oldest cohort (65 years old and older) to the cohort born between 1945 and 1949,[17]

---

[16] The numbers in brackets are the lower and upper 95% confidence intervals estimated using the bias-corrected percentile method and 500 resamplings for each case.

[17] The database is from 2004.

between-group inequality increased, and then younger generations experienced a steady decrease. According to a sub-population invariant perspective, i.e. using the Overlap index, such reduction in inequality was monotonic cohort-by-cohort. By contrast, the dissimilarity index, which takes into account changes in the composition of the groups' relative populations, reports an increase in inequality from the second-to-youngest to the youngest cohort. Yet even in that case, according to H, between-group educational inequality still remains below the levels corresponding to the cohorts of people 40 years old and older. Therefore, both indices provide evidence supporting the conclusion that the youngest generations of India are characterized by lower between-group inequality of educational attainment than their older peers. Since groups are defined by combinations of gender and caste origin, i.e. characteristics for which people can not be held accountable, these results reveal a partial reduction in inequality of educational opportunity.[18]

Table 1: Changes in educational opportunity across cohorts of Indian adults

Cohort	Dissimilarity index	Overlap index
30-34	0.155 [0.152, 0.159]	0.499 [0.483, 0.514]
35-39	0.149 [0.145, 0.152]	0.503 [0.486, 0.522]
40-44	0.163 [0.159, 0.167]	0.564 [0.543, 0.583]
45-49	0.170[0.165, 0.173]	0.591[0.572, 0.612]
50-54	0.185 [0.181, 0.190]	0.620[0.598, 0.641]
55-59	0.195 [0.190, 0.200]	0.657 [0.629, 0.680]
60-64	0.195 [0.189, 0.200]	0.650[0.628, 0.671]
65+	0.188 [0.183, 0.192]	0.622 [0.602, 0.636]

[18]Partial, because other characteristics beyond individuals' control, e.g. several dimensions of parental and family background, are not accounted for in this illustrative application.

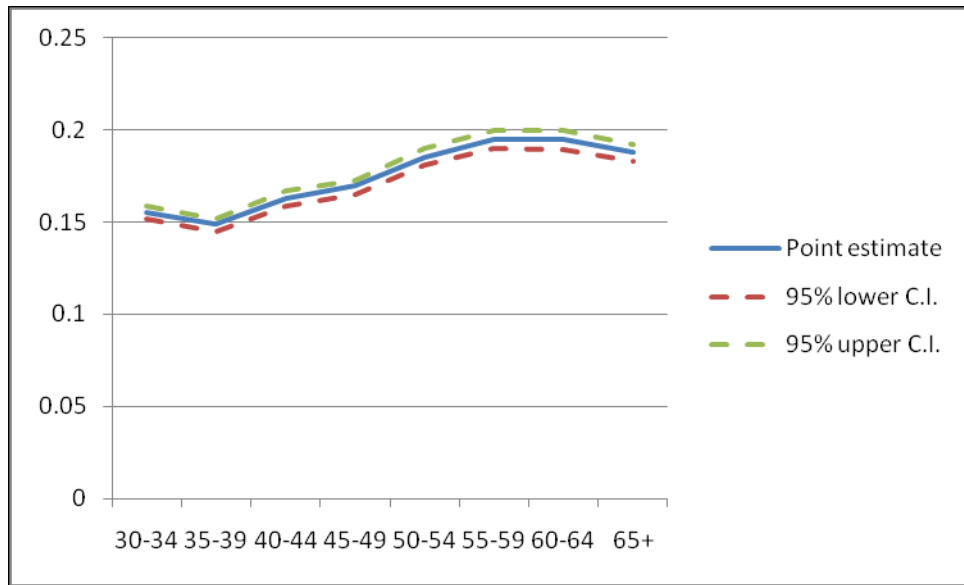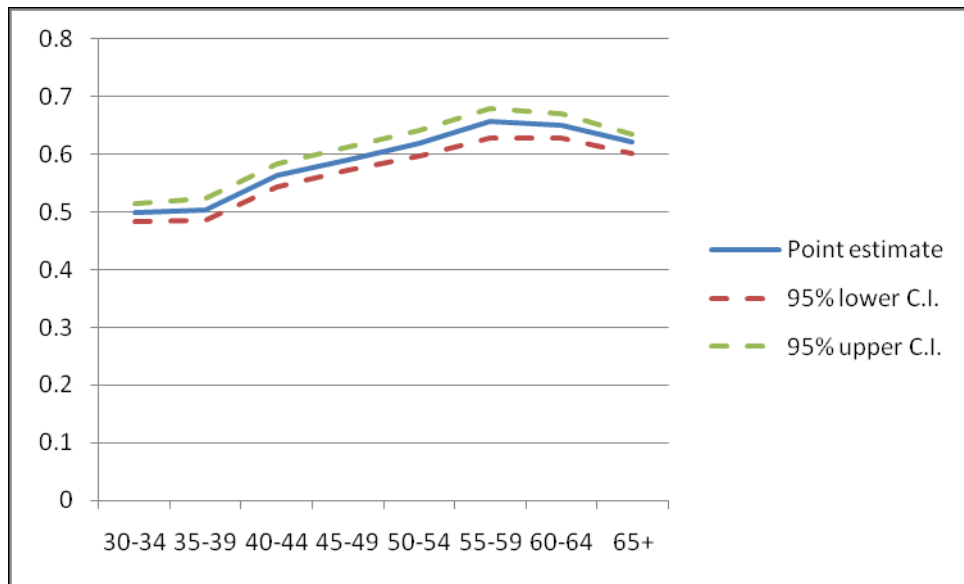Figure 1: Changes in educational opportunity across cohorts according to H

Figure 2: Changes in educational opportunity across cohorts according to O

6 Concluding remarks

The Pearson-based dissimilarity index and the multi-group overlap index are suitable for the comparison of multivariate and multinomial distributions of wellbeing across different groups. This paper shows that the two indices share two crucial traits: They attain a level of minimum between-group inequality if and only if group-specific distributions are all homogeneous, and they respond equally to minimum intra-group migrations (MIMs) that restore or break pairwise state between-group equality (PSBE). However, the paper shows that they also differ in three important aspects:

- They are sensitive (or insensitive) to MIMs that neither restore nor break PSBE in a priori differential ways. In fact in many circumstances O may be insensitive, or yield a null effect, to a MIM while H reacts to that same MIM, and viceversa. In general, H is more sensitive to MIMs than O.

- They fulfill different population invariance axioms. Chiefly, while H is sensitive to changes in group composition, O is not. The latter compares group's distributions without weighting them by group size thereby taking a "representative agent" approach to the comparison problem. This distinction provides a criterion for the researcher to choose between the two indices depending on his/her interest to account for (or, alternatively neutralize) the impact of varying group composition.

- They attain their maximum values in relation to slightly different concepts of maximum association. The paper proves that H attains its maximum if and only if there is complete association between groups and outcome states (or absolute association if $T = A$). By contrast, O attains its maximum if and only if a condition termed minimum overlap maximum association (MOMA) holds. Crucially, the paper shows that for every combination of T and A, $H = 1 \rightarrow O = 1$.

Studying the maxima of both indices the paper highlights their usefulness to test null hypotheses of complete or absolute association.

Finally, the paper uses the two indices to measure changes in between-group inequality of educational attainment in India across different cohorts of adults. Because the groups are defined in terms of gender and caste, such analysis documents differences in partial educational opportunity. Despite some differences in the indices' empirical assessment, which are explainable by their conceptual differences as discussed in the paper, they provide evidence in favour of a reduction of between-group inequality of educational attainment among the youngest cohorts of adult Indian citizens.

References

ANDERSON, G., Y. GE, AND T. W. LEO (2010): "Distributional overlap: simple, multivariate, parametric, and nonparametric tests for alienation, convergence, and general distribtuional difference issues," *Econometric reviews*, 29(3), 247–75.

ANDERSON, G., O. LINTON, AND Y.-J. WHANG (2009): "Nonparametric estimation of a polarization measure," manuscript.

ANDERSON, T., AND L. GOODMAN (1957): "Statistical inference about Markov chains," *The annals of Mathematical Statistics*, 28(1), 89–110.

BARROS, R. P. D., F. FERREIRA, J. MOLINAS, AND J. SAAVEDRA (2009): *Measuring inequality of opportunity in Latin America and the Caribbean*. The World Bank.

BRETON, M. L., A. MICHELANGELI, AND E. PELUSO (2008): "Wage discrimination measurement: in defense of a simple but informative statistical tool," manuscript.

CHANDRA, B., M. MUKHERJEE, AND A. MUKHERJEE (2008): *India since independence*. Penguin Books.

CHECCI, D., AND V. PERAGINE (2005): "Regional disparities and inequality of opportunity: the case of Italy," Discussion Paper Series, 1874, IZA.

DAGUM, C. (1987): "Measuring the economic affluence between populations of income receivers," *Journal of Business and Economic Statistics*, 5(1), 5–8.

DESHPANDE, A. (2007): "Overlapping identities under liberalization: Gender and Caste in India," *Economic Development and Cultural Change*, pp. 735–60.

DREZE, J., AND A. SEN (1995): *India: Economic Development and Social Opportunity*. Oxford University Press.

DUCLOS, J.-Y., J. ESTEBAN, AND D. RAY (2004): "Polarization: concepts, measurement, estimation," *Econometrica*, 72(6), 1737–72.

EBERT, U. (1984): "Measures of distance between income distributions," *Journal of Economic Theory*, 32, 266–74.

ELBERS, C., P. LANJOUW, J. MISTIAEN, AND B. OZLER (2008): "Reinterpreting between-group inequality," *Journal of Economic Inequality*, 6, 231–45.

ESTEBAN, J.-M., AND D. RAY (1994): "On the measurement of polarization," *Econometrica*, 62(4), 819–51.

FLEURBAEY, M. (2008): *Fairness, Responsibility and Welfare*. Oxford University Press.

FOSTER, J., AND A. SHNEYEROV (2000): "Path independent inequality measures," *Journal of economic theory*, 91, 199–222.

HANDCOCK, M., AND M. MORRIS (1999): *Relative distribution methods in the social sciences*, Statistics for social science and public policy. Springer.

KENDALL, M., AND A. STUART (1973): *The Advanced Theory of Statistics*, vol. 2. Griffin, 3 edn.

OPHI, AND U. DE CHILE (2009): "Encuesta - Otras dimensiones de la calidad de vida en los hogares," Discussion paper, OPHI and Universidad de Chile.

13

ROEMER, J. (1998): *Equality of Opportunity*. Harvard University Press.

WEITZMAN, M. (1970): "Measures of overlap of income distributions of white and negro families in the U.S.," Technical Paper 22, Bureau of the census.

YALONETZKY, G. (2009): "A dissimilarity index of multidimensional inequality of opportunity," OPHI Working Paper 28.

———— (2010): "Measuring group disadvantage with indices based on relative distributions," OPHI Research-in-progress.

7 Appendix: Descriptive statistics

Table 2: Distributions of educational attainment by gender-caste groups: Adults 30-34 years old

Groups	Educational categories						
	1	2	3	4	5	6	7
Male scheduled tribe	0.211	0.033	0.102	0.159	0.195	0.111	0.188
Female scheduled tribe	0.407	0.033	0.101	0.132	0.179	0.072	0.076
Male scheduled caste	0.257	0.027	0.096	0.159	0.201	0.096	0.163
Female scheduled caste	0.566	0.022	0.084	0.119	0.104	0.056	0.048
Male other backward castes	0.167	0.021	0.079	0.153	0.219	0.135	0.227
Female other backward castes	0.440	0.022	0.077	0.133	0.150	0.084	0.093
Male other castes	0.106	0.017	0.055	0.114	0.175	0.176	0.357
Female other castes	0.253	0.024	0.075	0.131	0.166	0.125	0.225

Table 3: Distributions of educational attainment by gender-caste groups: Adults 35-39 years old

Groups	Educational categories						
	1	2	3	4	5	6	7
Male scheduled tribe	0.254	0.035	0.097	0.147	0.189	0.097	0.181
Female scheduled tribe	0.440	0.032	0.098	0.141	0.166	0.060	0.064
Male scheduled caste	0.336	0.028	0.110	0.160	0.162	0.087	0.117
Female scheduled caste	0.621	0.023	0.087	0.106	0.086	0.039	0.038
Male other backward castes	0.224	0.025	0.096	0.157	0.200	0.124	0.174
Female other backward castes	0.485	0.022	0.084	0.139	0.129	0.068	0.073
Male other castes	0.141	0.022	0.065	0.136	0.163	0.157	0.315
Female other castes	0.293	0.024	0.076	0.146	0.161	0.123	0.176

Table 4: Distributions of educational attainment by gender-caste groups: Adults 40-44 years old

Groups	Educational categories						
	1	2	3	4	5	6	7
Male scheduled tribe	0.270	0.040	0.110	0.144	0.168	0.089	0.179
Female scheduled tribe	0.501	0.034	0.111	0.121	0.126	0.048	0.059
Male scheduled caste	0.381	0.029	0.112	0.155	0.143	0.076	0.104
Female scheduled caste	0.702	0.021	0.069	0.094	0.059	0.031	0.024
Male other backward castes	0.253	0.028	0.105	0.160	0.180	0.111	0.164
Female other backward castes	0.533	0.021	0.088	0.136	0.115	0.058	0.050
Male other castes	0.144	0.019	0.065	0.134	0.179	0.150	0.309
Female other castes	0.335	0.027	0.080	0.144	0.145	0.113	0.156

Table 5: Distributions of educational attainment by gender-caste groups: Adults 45-49 years old

Groups	Educational categories						
	1	2	3	4	5	6	7
Male scheduled tribe	0.301	0.047	0.098	0.147	0.149	0.084	0.174
Female scheduled tribe	0.559	0.039	0.101	0.113	0.095	0.0387	0.054
Male scheduled caste	0.405	0.036	0.110	0.155	0.128	0.069	0.097
Female scheduled caste	0.743	0.027	0.056	0.081	0.054	0.021	0.018
Male other backward castes	0.264	0.035	0.114	0.150	0.176	0.110	0.152
Female other backward castes	0.575	0.028	0.086	0.122	0.100	0.049	0.040
Male other castes	0.155	0.024	0.069	0.130	0.177	0.152	0.293
Female other castes	0.381	0.026	0.077	0.133	0.141	0.105	0.137

Table 6: Distributions of educational attainment by gender-caste groups: Adults 50-54 years old

Groups	Educational categories						
	1	2	3	4	5	6	7
Male scheduled tribe	0.343	0.044	0.110	0.160	0.131	0.076	0.137
Female scheduled tribe	0.642	0.034	0.098	0.111	0.054	0.029	0.032
Male scheduled caste	0.427	0.034	0.115	0.133	0.125	0.075	0.093
Female scheduled caste	0.774	0.016	0.065	0.069	0.042	0.011	0.022
Male other backward castes	0.304	0.033	0.097	0.145	0.155	0.113	0.153
Female other backward castes	0.661	0.019	0.078	0.095	0.072	0.038	0.036
Male other castes	0.154	0.024	0.070	0.120	0.162	0.160	0.310
Female other castes	0.423	0.031	0.089	0.124	0.119	0.092	0.122

Table 7: Distributions of educational attainment by gender-caste groups: Adults 55-59 years old

Groups	Educational categories						
	1	2	3	4	5	6	7
Male scheduled tribe	0.395	0.069	0.119	0.147	0.110	0.061	0.098
Female scheduled tribe	0.746	0.031	0.070	0.085	0.042	0.006	0.020
Male scheduled caste	0.481	0.039	0.100	0.119	0.096	0.077	0.089
Female scheduled caste	0.846	0.016	0.045	0.044	0.026	0.010	0.013
Male other backward castes	0.326	0.035	0.132	0.154	0.120	0.097	0.136
Female other backward castes	0.713	0.019	0.080	0.092	0.048	0.023	0.025
Male other castes	0.193	0.029	0.083	0.128	0.142	0.150	0.275
Female other castes	0.488	0.033	0.087	0.136	0.106	0.068	0.082

Table 8: Distributions of educational attainment by gender-caste groups: Adults 60-64 years old

Groups	Educational categories						
	1	2	3	4	5	6	7
Male scheduled tribe	0.554	0.053	0.107	0.127	0.068	0.042	0.049
Female scheduled tribe	0.810	0.028	0.064	0.051	0.024	0.007	0.015
Male scheduled caste	0.593	0.037	0.102	0.095	0.073	0.054	0.045
Female scheduled caste	0.907	0.015	0.028	0.026	0.011	0.007	0.007
Male other backward castes	0.435	0.032	0.117	0.149	0.102	0.083	0.081
Female other backward castes	0.796	0.018	0.065	0.064	0.028	0.016	0.012
Male other castes	0.257	0.029	0.083	0.138	0.127	0.154	0.212
Female other castes	0.591	0.027	0.082	0.125	0.071	0.056	0.048

Table 9: Distributions of educational attainment by gender-caste groups: Adults 65+ years old

Groups	Educational categories						
	1	2	3	4	5	6	7
Male scheduled tribe	0.614	0.055	0.107	0.090	0.077	0.027	0.031
Female scheduled tribe	0.834	0.025	0.069	0.043	0.019	0.007	0.003
Male scheduled caste	0.681	0.038	0.094	0.085	0.051	0.024	0.028
Female scheduled caste	0.936	0.012	0.023	0.016	0.009	0.002	0.002
Male other backward castes	0.500	0.045	0.135	0.147	0.067	0.057	0.049
Female other backward castes	0.836	0.019	0.061	0.054	0.019	0.007	0.005
Male other castes	0.314	0.043	0.115	0.138	0.125	0.131	0.135
Female other castes	0.664	0.033	0.091	0.099	0.053	0.033	0.027

Table 10: Sample sizes by group and cohort

Groups	Cohorts							
	30-34	**35-39**	**40-44**	**45-49**	**50-54**	**55-59**	**60-64**	**65+**
Male scheduled tribe	2831	2890	2416	2224	1570	1209	876	1137
Female scheduled tribe	3159	2885	2207	1914	1418	1098	822	1180
Male scheduled caste	3354	3498	2852	2493	1849	1424	1122	1935
Female scheduled caste	3706	3435	2629	2277	1656	1367	1308	1961
Male other backward castes	8150	8040	6649	6103	4613	3612	3022	5208
Female other backward castes	8789	8273	6425	5500	4361	3696	3165	5498
Male other castes	7548	7316	6539	5891	4482	3628	2917	5550
Female other castes	7762	7364	6067	5291	4118	3345	2989	5466