UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper
Economic Series 10-15
June 2010

Departamento de Economía
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 916249875

A KULLBACK-LEIBLER MEASURE OF CONDITIONAL SEGREGATION

Ricardo Mora and Javier Ruiz-Castillo[1]
Departamento de Economía, Universidad Carlos III de Madrid

Abstract

In this paper the Kullback-Leibler notion of discrepancy (Kullback and Leibler, 1951) is used to propose a measure of multigroup segregation over a set of organizational units within a multivariate framework. Among the main results of the paper it is established that the Mutual Information index of segregation, *M*, first proposed by Theil and Finizza (1971), whose ranking has been fully characterized in terms of seven ordinal axioms by Frankel and Volij (2009), can be decomposed to isolate a term which captures segregation conditional on any vector of covariates. Furthermore, consistent estimators for *M* and the terms in its decomposition are proposed, and their asymptotic properties are obtained. The usefulness of the approach is illustrated by looking at patterns of multiracial segregation across public schools in the U.S. for the academic years 1989-90 and 2005-06. It is found that most within-cities segregation and a significant part of within-districts segregation is accounted for by county-level income per capita and wages per job, and teachers per pupil at school level.

Keywords: Kullback-Leibler Discrepancy; Conditional Segregation; Asymptotic Properties; Econometric Models.

[1] This is a completely new version of a 2007 Working Paper entitled "The Statistical Properties of the Mutual Information Index of Multigroup Segregation". The authors acknowledge financial support from the Spanish DGI, Grants ECO2009-11165 and SEJ2007-67436.

I. INTRODUCTION

Social scientists have long been interested in the measurement of occupational segregation by gender, as well as in residential and educational segregation by ethnic group.[2] Mathematically, these problems are similar in the sense that both involve summarizing by means of a real number the information contained in the frequency of individuals (workers, residents, students) over a finite set of organizational units (occupations, neighbourhoods, schools) and a finite set of demographic groups (defined in terms of gender, racial, or ethnic categories). Such a real number is referred to as an index of segregation. For concreteness, this paper will use the example of school segregation in the multiracial case. The main question we address is how to account for racial group and school differences in socioeconomic variables in the measurement of segregation.

To place this issue into a proper perspective, think of school segregation at a national level as arising from two forces. Firstly, given the partition of cities into school districts, school segregation arises from politically determined segregative or integrative rules in the assignment of students to schools within a given district (see *inter alia* Rivkin, 1994, and Clotfelter, 1999). Secondly, imagine a situation without within-districts school segregation, that is, a situation where school district authorities all over the country are able to implement a policy that reproduces in all schools the racial mix of the district to which they belong. In this scenario, the student population would still experience some segregation arising from the residential choices adopted by their parents or caretakers: as long as the racial composition at the school district level differs from the racial composition at the city and/or the national level, there will be between-cities and between-districts (or within-cities) school segregation in the country as a whole. Preferences and opportunities behind residential decisions may directly depend on a number of socioeconomic variables, giving rise to the main issue addressed in this paper. Assume, for instance, that there is a statistical association between student race and household income levels. In

[2] For a treatise on occupational segregation by gender, see Fluckiber and Silber (1999), and for a recent useful contribution on residential and school segregation, see Reardon and Firebaugh (2002).

so far as household income is a potential determinant of residential and school choice, it can be said that multigroup school segregation may be partially due to income inequality. Therefore, for both explanatory and policy reasons it is important to identify the extent to which the value of segregation arises from income and other socioeconomic characteristics.

In the absence of a better strategy, one can discretize the vector of socioeconomic controls and use indices of segregation which are additively decomposable into between and within discrete categories, such as in Reardon *et al*. (2000), Mora and Ruiz-Castillo (2003), and Frankel and Volij (2009). However, this strategy has a practical limitation and a conceptual drawback. The practical limitation stems from the curse of dimensionality: to avoid serious aggregation bias, one should consider as many categories as possible for each control, but with usual sample sizes this is implementable in practice only when the vector of controls has few dimensions. The conceptual drawback is due to the absence of a clear interpretation of the between term as the used discrete categories are only arguable approximations of the actual values. Therefore, the use of indices which are additively decomposable into between and within discrete categories only partially answers this question because it does not deal properly with many continuous controls.

Other researchers have tried to develop notions of conditional segregation which should be implementable in a general multivariate framework. Frequently, their ultimate purpose is to assess to what extent segregation can be explained by the determinants of individual choice; to do so, they borrow the tools used in the literature on discrete choice. For example, in their analysis of occupational segregation, Spriggs and Williams (1996) propose a modified Duncan dissimilarity index which uses gender (and race) differences in estimated probabilities of being in an occupation obtained from multinomial logit models. Following closely Carrington and Troske (1997), other researchers have proposed indices of segregation which attempt to control for systematic differences in the distribution of covariates across groups. For example, in the context of occupational segregation by immigrant status, Aslund and Skans (2009) propose estimating the propensity score for each group given the vector of characteristics to create the benchmark random allocation (conditional on the covariates) for any segregation index.[3] They then develop a test of conditional segregation using an index of exposure. The most important drawback of these strategies is that the indices obtained are neither characterized in terms of axiomatic properties, nor related in an unambiguous way to indices which are fully characterized. This implies that although the procedures suggested sometimes have a clear intuitive appeal, it is not clear how they relate to unconditional measures of segregation and one cannot be certain of what the resulting index actually measures.

In this paper, a multivariate statistical framework to analyse multigroup school segregation is set up by borrowing the Kullback and Leibler (1951) notion of discrepancy from Information Theory. A measure of segregation, M, is then proposed and shown to satisfy several important properties.

Firstly, M coincides with the Mutual Information index, first proposed by Theil and Finizza (1971) as a measure of racial school segregation at district level, and whose ranking has been recently characterized by Frankel and Volij (2009) in terms of seven ordinal desirable axioms.

Secondly, Frankel and Volij (2009) show that, for any variable d which partitions the set of schools or the set of racial categories, M is strongly decomposable and the within term in this decomposition can be interpreted as segregation conditional on d. In this paper, this result is generalized to condition segregation on any vector of (possibly continuous) student and school characteristics $\mathbf{x}$. In particular, the M index can be decomposed into a between term, M_{KL}^{B}, which is a Kullback-Leibler measure of discrepancy and captures the statistical dependence between race status (or school membership) and $\mathbf{x}$, and a within term, M_{KL}^{W}, which captures multigroup school segregation conditional on $\mathbf{x}$. Because M_{KL}^{B} and M_{KL}^{W} are independent, in the sense that it is possible to introduce changes in the population to eliminate conditional segregation M_{KL}^{W} keeping conditioning segregation M_{KL}^{B} constant, this

[3] See also Hellerstein and Neumark (2008) and Kalter (2000) for related methodological proposals.

decomposition allows us to answer questions such as "to what extent is racial segregation at school level associated with racial differences in socioeconomic variables?"[4] Moreover, since M^B_{KL} and M^W_{KL} are functions of terms that can be interpreted as qualitative response models, the decomposition provides an intuitive unifying econometric framework for studies of segregation using segregation indices and econometric models.

Thirdly, since segregation measures are routinely computed using samples, it is usually of interest to study their statistical significance. The simplest approach to this problem involves reporting *t*-statistics using computer intensive methods such as the bootstrap as in Boisso *et al.* (1994). A related approach consists of standardizing the segregation measure, using as mean and standard deviations estimates obtained from resampling under random assignment into groups and organizational units, as in Carrington and Troske (1997). Other authors have made use of a statistical framework for the empirical analysis of segregation, as in Kakwani (1994). In this paper, for any sample of size T, estimators for both the M index, $\hat{M}_T$, and also the between and within terms in its decomposition, $\hat{M}^B_T$ and $\hat{M}^W_T$, are proposed using the principle of analogy. $\hat{M}_T$ is shown to be a monotonic transformation of the likelihood-ratio statistic for testing statistical independence between school membership and racial status. Furthermore, when the vector of covariates $\mathbf{x}$ only includes discrete variables, it is shown that $\hat{M}^W_T$ can be interpreted as a monotone transformation of the likelihood-ratio statistic for testing statistical independence between school membership and racial status given $\mathbf{x}$. Finally, sufficient conditions are provided to obtain under all segregation scenarios the asymptotic properties of $\hat{M}_T$, $\hat{M}^B_T$, and $\hat{M}^W_T$, both in the case when all variables are discrete and also when there is at least one continuous variable in $\mathbf{x}$.

[4] In the field of income inequality, between-groups income inequality can also be interpreted as the amount by which overall income inequality is reduced when the differences between subgroup income means are eliminated by making them equal to the population income mean (see, *inter alia*, Shorrocks, 1984). As shown by Mora and Ruiz-Castillo (2009), the corresponding interpretation is logically impossible in segregation studies.

To summarize, it has been shown elsewhere that M is well grounded on an axiomatic notion of segregation. In this paper, we show that it can be used to estimate the level of segregation which does not arise from the statistical association between the demographic groups and any set of covariates. The usefulness of the approach is illustrated by applying it to the analysis of multiracial segregation in the U.S. public schools. More specifically, we study to what extent the measures of within-cities and within-districts segregation are due to the statistical association between racial group membership and three continuous variables: county income per capita and wages per job, and teachers per pupil at district and school level. Results show that around 64% and 20% of, respectively, within-cities and within-districts segregation is accounted for by these three covariates, and that these shares are strongly significant.

The rest of the paper contains four sections. Section 2 sets up the general statistical framework, and defines M and its decomposition in a multivariate framework. Section 3 proposes estimators $\hat{M}_T$, $\hat{M}_T^B$, and $\hat{M}_T^W$ and presents the asymptotic results. Section 4 contains the empirical illustration, while Section 5 offers some concluding comments.

II. A GENERAL STATISTICAL MODEL OF MULTIGROUP SCHOOL SEGREGATION

II. 1. Measures of Segregation

It is useful to refer to a specific segregation problem. For consistency with the empirical illustration in Section IV, the case discussed throughout the paper is the multigroup school segregation problem. Assume a city **X** consisting of N schools, indexed by $n = 1,\ldots, N$. Each student belongs to any of G racial groups, indexed by $g = 1,\ldots, G$. The data available can be organized into the following G x N matrix:

$$\mathbf{X} = \{t_{gn}\} = \begin{bmatrix} t_{11} & \cdots & t_{1N} \\ \vdots & \ddots & \vdots \\ t_{G1} & \cdots & t_{GN} \end{bmatrix}, \tag{1}$$

where t_{gn} is the number of individuals of racial group g attending school n, so that $t=\sum_{n=1}^{N}\sum_{g=1}^{G} t_{gn}$ is the total student population.

The information contained in the joint absolute frequencies of racial groups and schools, t_{gn}, is usually summarized by means of numerical indices of segregation. Let $X(G, N)$ be the set of all cities with G groups and N schools. A segregation index S is a real valued function defined in $X(G, N)$, where $S(\mathbf{X})$ provides the extent of school segregation for any city $\mathbf{X}\in X(G, N)$. Let $p_{gn} = t_{gn}/t$, and denote by $P_{gn}=\left\{p_{gn}\right\}_{g=1,n=1}^{G,N}$, the joint distribution of racial groups and schools in a city $\mathbf{X}\in X(G, N)$. In the following section, the discussion will be restricted to indices that capture a *relative* view of segregation in which all that matters is the joint distribution, i.e. indices which admit a representation as a function of P_{gn}.[5]

II. 2. A Kullback-Leibler Measure of Segregation

Consider the probability space $(\Omega, \mathcal{F}, m)$ where Ω is the set of possible samples $\{g, n, \mathbf{x}\}\in\Omega$ where $\mathbf{x}\in\Lambda\subset\mathbb{R}^k$ is a vector of k covariates. $\mathcal{F}$ is the σ-algebra of subsets of Ω, and m is a measure of the probability of the events in $\mathcal{F}$. Assume that there are two absolutely continuous measures with respect to m, m_1 and m_2, and two generalized density functions, $f_1(g, n, \mathbf{x})$ and $f_2(g, n, \mathbf{x})$, such that

$$m_i(E)=\int_E f_i(g,n,\mathbf{x})dm,\ i=1,\,2,$$

for all $E\in\mathcal{F}$. The elements in $\mathbf{x}$ may be univariate or multivariate, discrete or continuous, qualitative or quantitative, and the generalized density functions f_i are known at most up to a parameter vector. Consider the partition of Ω into G x N sets $D_{gn}=\{(r,s,\mathbf{x})\in\mathcal{F}\text{: } r=g,\ s=n, \mathbf{x}\in\Lambda\}$ and let

[5] This property, satisfied by most segregation indices, is referred to as *Size Invariance* in James and Taeuber (1985).

$$\mu_i(g,n) = \int_{D_{gn}} f_i(r,s,\mathbf{x})d\mu,\ i=1,2,$$

so that the probability that a student is of race g and belongs to school n under the probability measure μ_1 is $p_{gn} = \mu_1(g,n) = \int_{D_{gn}} f_1(r,s,\mathbf{x})d\mu$, where $p_{gn} \geq 0$ and $\sum_{g=1}^{G}\sum_{n=1}^{N} p_{gn} = 1$. The marginal probabilities for race status and school membership are $p_{g\bullet} = \sum_{n=1}^{N} p_{gn}$ and $p_{\bullet n} = \sum_{g=1}^{G} p_{gn}$, respectively. For all g and n such that $\mu_i(g,n) > 0,\ i=1,\ 2,$ the generalized conditional density given race and school status is $f_i(\mathbf{x} \mid g,n) = \dfrac{f_i(g,n,\mathbf{x})}{\mu_i(g,n)}$. Following Kullback (1959), a Kullback-Leibler, *KL*, measure of discrepancy between f_1 and f_2 is defined as:

$$I(1:2) = \int f_1(g,n,\mathbf{x})\log\left(\frac{f_1(g,n,\mathbf{x})}{f_2(g,n,\mathbf{x})}\right)d\mu. \tag{1}$$

Let H_i , i =1, 2, represent the hypothesis that $(g,n,\mathbf{x})$ belongs to the statistical population with probability measure μ_i, and define the logarithm of the likelihood ratio, $\log\left(\dfrac{f_1(g,n,\mathbf{x})}{f_2(g,n,\mathbf{x})}\right)$, as the information in $(g,n,\mathbf{x})$ for discrimination in favour of H_1 against H_2.[6] Then $I(1:2)$ can be interpreted as the mean discrepancy (or information for discrimination) in favour of H_1 against H_2 per observation from μ_1 (see Kullback, 1959, p. 5).

Define the conditional probability of school membership n given race status g as $p_{n|g} = \dfrac{p_{gn}}{p_{g\bullet}}$, and let $P_{n|g} = \left\{p_{n|g}\right\}_{n=1}^{N}$ represent the conditional distribution of students from group g across schools.

[6] The base of the logarithm is immaterial, providing essentially a unit of measure. The natural logarithm is used throughout the paper.

Similarly, define the conditional probability of racial status g given school membership n as $p_{g|n} = \dfrac{p_{gn}}{p_{\bullet n}}$, and denote by $P_{g|n} = \left\{ p_{g|n} \right\}_{g=1}^{G}$ the racial mix within school n. Indices in the segregation literature associate the absence of segregation with two situations. Firstly, racial groups are not segregated if the relative frequency with which a student attends school n is constant, regardless of her racial group, i.e. $p_{n|g} = p_{\bullet n}$.[7] Secondly, the racial composition at all schools is fully representative of the population if the relative frequency with which students belong to racial group g is constant regardless of the school which they attend, i.e. $p_{g|n} = p_{g\bullet}$.[8] These two notions of absence of segregation are equivalent and coincide with the concept of statistical independence between race status and school membership:

$$p_{g|n} = p_{g\bullet} \Leftrightarrow p_{n|g} = p_{\bullet n} \Leftrightarrow p_{ng} = p_{g\bullet} p_{\bullet n}.$$

Under the following three assumptions the KL notion of discrepancy between dependence and independence of race and school membership becomes a measure of segregation. For all $g = 1,\ldots, G$, $n = 1,\ldots, N$, and $\mathbf{x} \in \Lambda \subset \mathbb{R}^k$:

A1 : $p_{gn} > 0$.

A2 : $f_i(\mathbf{x} \mid g,n) = f(\mathbf{x} \mid g,n) > 0$ as, $i = 1,2$.

A3 : $m_2(g,n) = p_{g\bullet} p_{\bullet n} = \left(\sum_{n=1}^{N} p_{gn} \right) \left(\sum_{g=1}^{G} p_{gn} \right)$.

A1 eliminates from consideration combinations of races and schools that are *a priori* impossible to observe. A2 ensures that the marginal probabilities p_{gn} are sufficient statistics with respect to the measure of discrepancy, so that no information is lost by disregarding $\mathbf{x}$. Finally, A3 identifies H_2 with the notion

[7] Absence of segregation in this sense is consistent with the notion of segregation as "evenness", advocated by James and Taeuber (1985), according to which segregation is seen as the tendency of racial groups to have different distributions across schools.

[8] Absence of segregation in this sense follows the idea of "representativeness", emphasized by Frankel and Volij (2009), which asks to what extent schools have different racial compositions from the population as a whole, and it is closely related to the idea of "isolation" distinguished by Massey and Denton (1988) in the two-group case.

of statistical independence between race and school membership. Given equation (1), the following remark results.

Remark 1: Under assumptions A1 to A3, the notion of discrepancy $I(1:2)$ coincides with the Mutual Information index, *M*, i.e.

$$I(1:2)=\sum_{g=1}^{G}\sum_{n=1}^{N}p_{gn}\log\left(\frac{p_{gn}}{p_{g\bullet}p_{\bullet n}}\right)=\sum_{g=1}^{G}p_{g\bullet}\sum_{n=1}^{N}p_{n|g}\log\left(\frac{p_{n|g}}{p_{\bullet n}}\right)=M.$$

Theil (1972) shows that *M* is bounded. The lower bound 0 is achieved whenever $p_{ng}=p_{g\bullet}p_{\bullet n}$ for all *g* and *n*, while the upper bound is $\min\{\log(G),\ \log(N)\}$.

II. 2. Multigroup School Conditional Segregation

Assumptions A1 to A3 do not require independence between race status (or school membership) and any of the covariates in **x**. Thus, as is pointed out in the Introduction, it will be generally of interest to evaluate the extent to which *M* can be attributed to the statistical association between the covariates **x** and the racial groups (or schools). Without loss of generality, let us consider the statistical association between racial groups and covariates **x**.

It is always possible to factorize the generalized density $f_i(g,n,\mathbf{x})$ as $f_i(g,n,\mathbf{x})=f_i(n\,|\,g,\mathbf{x})f_i(g,\mathbf{x})$, where $f_i(g,\mathbf{x})=\sum_{n=1}^{N}f_i(g,n,\mathbf{x})$, $i=1,\ 2$. Therefore, any measure of discrepancy $I(1:2)$ can always be decomposed into two terms:

$$\begin{aligned}I(1:2)=&\int f_1(g,n,\mathbf{x})\log\left(\frac{f_1(g,\mathbf{x})}{f_2(g,\mathbf{x})}\right)dm\\&+\int f_1(g,n,\mathbf{x})\log\left(\frac{f_1(n\,|\,g,\mathbf{x})}{f_2(n\,|\,g,\mathbf{x})}\right)dm.\end{aligned}\qquad(2)$$

The first term captures the discrepancy between $f_1(g,\mathbf{x})$ and $f_2(g,\mathbf{x})$, while the second term captures the discrepancy in conditional school assignment rules $f_1(n\,|\,g,\mathbf{x})$ and $f_2(n\,|\,g,\mathbf{x})$. In addition to A1, A2, and A3, the following four assumptions are sufficient to obtain a decomposition of *M* so that one term can be interpreted as racial discrepancy across covariates and the other term can be interpreted as conditional school segregation:

A4 : $f_1(g,\mathbf{x}) = f(g\,|\,\mathbf{x};a)f(\mathbf{x})$ where $f(\bullet\,|\,\mathbf{x};a)$ is known up to parameter vector $a \in \mathbb{R}^{k_a}$.

A5 : $f_2(g,\mathbf{x}) = p_{g\bullet}f(\mathbf{x})$ with $p_{g\bullet} = \int_{\mathbf{x}\in\Lambda} f(g\,|\,\mathbf{x};a)f(\mathbf{x})d\mathbf{x}$ not uniquely identified by a.

A6 : $f_1(g,n\,|\,\mathbf{x}) = f(g,n\,|\,\mathbf{x};b)$ where $f(\bullet,\bullet\,|\,\mathbf{x};b)$ is known up to parameter vector $b \in \mathbb{R}^{k_b}$ which is not a function of (g,a) where $g = (p_{1\bullet},\ldots,p_{G\bullet})'$.

A7 : $f_2(g,n\,|\,\mathbf{x}) = f(g\,|\,\mathbf{x};a)f(n\,|\,\mathbf{x};b)$ where $f(n\,|\,\mathbf{x};b) = \left\{\sum_{g=1}^{G} f(g,n\,|\,\mathbf{x};b)\right\}$.

Under assumptions A1 to A7, the M index can be decomposed into a between term which captures the statistical dependence between race status (or school membership) and **x**, and a within term which captures multigroup school segregation conditional on **x** (see Proposition 1 in Appendix):

$$M = M^B(g,a) + M^W(a,b) \tag{3}$$

where

$$M^B(g,a) = \int_{\mathbf{x}\in\Lambda} f(\mathbf{x})\left\{\sum_{g=1}^{G} f(g\,|\,\mathbf{x};a)\log\left(\frac{f(g\,|\,\mathbf{x};a)}{p_{g\bullet}}\right)\right\}d\mathbf{x}$$

and

$$M^W(a,b) = \int_{\mathbf{x}\in\Lambda} f(\mathbf{x})\left\{\sum_{g=1}^{G}\sum_{n=1}^{N} f(g,n\,|\,\mathbf{x};b)\log\left(\frac{f(g,n\,|\,\mathbf{x};b)}{f(g\,|\,\mathbf{x};a)f(n\,|\,\mathbf{x};b)}\right)\right\}d\mathbf{x}.$$

The term $M^B(g,a)$ identifies the level of segregation which would remain if there were no segregation after controlling for the statistical dependence between the vector of covariates **x** and racial status. Since $M^W(a,b)$ is the level of segregation which is not related to racial discrepancy by covariates **x**, it can be referred to as school segregation by race conditional on **x**.

Decomposition (3) is appealing for at least two reasons. Firstly, $M^W(a,b)$ and $M^B(g,a)$ are independent in the sense that it is possible to introduce changes in the densities to eliminate $M^W(a, b)$ keeping $M^B(g, a)$ constant. Secondly, conditional densities $f(n \mid \mathbf{x}; b)$, and $f(n \mid g, \mathbf{x}; b) \equiv f(g, n \mid \mathbf{x}; b) / f(g \mid \mathbf{x}; b)$ can be interpreted as qualitative response models which stem from economic agents' utility maximizing choices under constraints. Thus, decomposition (3) provides an intuitive, unifying, econometric framework for studies of segregation using segregation indices and qualitative response econometric models.

Kullback (1959) points out that any *KL* discrepancy can be recursively decomposed into more than two terms. This is trivially seen with decomposition (3), as the first term is itself a *KL* discrepancy measure and, hence, it can itself be decomposed. A direct application of this property to the problem of multigroup school segregation permits the decomposition of M into three terms capturing between-cities, within-cities, and within-districts school segregation.[9] For reasons of brevity, we leave to the reader the details of decompositions of more than two terms in the model.

One final point needs to be clarified. Suppose that all covariates $\mathbf{x}$ are discrete, and that they partition the set of schools into disjoint subsets, such as when schools in a city are organized into a set of school districts. More specifically, assume that each school belongs to one of K different school districts and let p_{gnd} denote the proportion of students of racial group g at school n within district d, $p_{gnd} = p_{gn|n \in d}$. Define $p_{g \bullet d}$ as the joint probability of race and district membership and let $p_{\bullet\bullet d}$ and $p_{gn|d}$ denote the marginal distribution of districts and the joint distribution of race and school membership conditional on district d, respectively. Finally, let $p_{g\bullet|d}$ and $p_{\bullet n|d}$ be the marginal distributions within district d of race and school membership. It has previously been shown that the M index is decomposable for any partition of the schools into K school districts into a between and a within term:[10]

[9] See also Hernanz *et al*. (2005) for an application of this principle in the context of occupational segregation by gender and Frankel and Volij (2009) for sequential clustering of racial categories in multiracial school segregation.
[10] See Frankel and Volij (2009) and Mora and Ruiz-Castillo (2009). For the two-group case, see Mora and Ruiz-Castillo (2003, 2004), and Herranz *et al*. (2005).

$$M = M^B + M^W, \tag{4}$$

where $M^B = \sum_{d=1}^{K}\sum_{g=1}^{G} p_{g\bullet d}\log\left(\frac{p_{g\bullet d}}{p_{g\bullet}\,p_{\bullet\bullet d}}\right)$ and $M^W = \sum_{k=1}^{K} p_{\bullet\bullet d}\sum_{n\in d}\sum_{g=1}^{G} p_{gn|d}\log\left(\frac{p_{gn|d}}{p_{g\bullet|d}\,p_{\bullet n|d}}\right)$.

How does decomposition (4) relate to decomposition (3)? If $f(\mathbf{x}) = p_{\bullet\bullet d}$, $f(g\,|\,\mathbf{x}, a) = p_{g\bullet|d}$, $f(g,n\,|\,\mathbf{x}; b) = p_{gn|d}$, and $f(n\,|\,\mathbf{x}; b) = p_{\bullet n|d}$, it can readily be shown that $M^B(g, a) = M^B$ and $M^W(a, b) = M^W$. Thus, when the vector of covariates includes only discrete variables, the general decomposition in equation (3) exactly matches the decomposition of the M index previously proposed in the literature for any partition of schools (or groups) as in equation (4).

III. ESTIMATION AND ASYMPTOTIC PROPERTIES

III.1. Estimation and Asymptotics of the M Index

Assume that a sample of T observations from students with information on their race status, school membership, and covariates $\mathbf{x}_i$, $\{g_i, n_i, \mathbf{x}_i\}_{i=1}^{T}$, is available. Let T_{gn} be the number of students of racial group g in school n, so that $T = \sum_{n=1}^{N}\sum_{g=1}^{G} T_{gn}$. Let $T_{g\bullet} = \sum_{n=1}^{N} T_{gn} > 0$ and $T_{\bullet n} = \sum_{g=1}^{G} T_{gn} > 0$. Note that under assumptions A1 to A7, race and school status are jointly distributed as a nonparametric multinomial model. Without loss of generality, denote by $GN^c = \{(g,n) : g = 1,...,G, n = 1,...,N, (g,n) \neq (G,N)\}$ the set of all race and school combinations except combination (G,N). Then the marginal probabilities of race and school membership are fully identified by the vector $q = (p_{11}, p_{21}, ..., p_{G,N-1})' \in \Theta \subset \mathbb{R}^{GN-1}$, with $\Theta \equiv \left\{\{p_{gn}\}_{GN^c}, \sum_{GN^c} p_{gn} < 1, p_{gn} > 0\right\}$, and $p_{GN} = 1 - \sum_{(g,n)\in GN^c} p_{gn} > 0$. The M index is bounded and continuous. Moreover, it can always be estimated by $\hat{M}_T = \sum_{g=1}^{G}\sum_{n=1}^{N} \hat{p}_{gn} \log\left(\frac{\hat{p}_{gn}}{\hat{p}_{g\bullet}\hat{p}_{\bullet n}}\right)$, where $\hat{p}_{gn} = T_{gn}/T$, $\hat{p}_{g\bullet} = T_{g\bullet}/T$, $\hat{p}_{\bullet n} = T_{\bullet n}/T$, and $0\log(0) = 0$.

As is shown in Proposition 2 in the Appendix, under assumptions A1 to A3, $\text{plim}\ \hat{M}_T = M$. An implication of this consistency result is that $\hat{M}_T$ converges in probability to 0 if and only if $p_{gn} = p_{g\bullet}p_{\bullet n}$ for all g and n. Moreover, whenever two cities are to be ranked according to *M*, the implicit ordering from $\hat{M}_T$ convergences in probability to the ordering induced by *M*.

The relation between Kullback-Leibler discrepancy measures and likelihood-ratio statistics for testing independence across categorical variables in contingency tables is well known (Kullback, 1959, p. 158). Here, to implement the likelihood ratio statistic for the independence of racial status and school membership, an additional parametric assumption on the conditional density for covariates **x** is sufficient:

<u>A8</u> : $f(\mathbf{x} \mid g,n) = f(\mathbf{x} \mid g,n;j)$ such that $f(\mathbf{x}) = \sum_{g=1}^{G}\sum_{n=1}^{N} f(\mathbf{x} \mid g,n;j)p_{gn}$ and $j \in \mathbb{R}^{k_j}$ does not depend on q.

Consider testing for the independence of race and school membership, i.e. $H_0 : p_{gn} = p_{g\bullet}p_{\bullet n}$ for all $(g,n) \in GN^c$, versus $H_1 : p_{gn} \neq p_{g\bullet}p_{\bullet n}$. Let $l(\hat{q},\hat{j})$ be the log-likelihood evaluated at the maximum likelihood (hereafter ML) estimator, and let $l(\hat{q}_0,\hat{j}_0)$ be the log-likelihood for the model under H_0 evaluated at the restricted ML estimator, so that $-2\log(l) = -2\left(l(\hat{q}_0,\hat{j}_0) - l(\hat{q},\hat{j})\right)$ is the log-likelihood ratio statistic and l is the likelihood ratio.

Remark 2: Under assumptions A1, A2, A3 and A8, $\hat{M}_T = \frac{-\log(l)}{T}$.

Therefore, $\hat{M}_T$ is a monotonic transformation of the likelihood-ratio statistic for testing statistical independence between school membership and racial status. This implies that the ordering across cities provided by comparisons of city-specific log-likelihood ratios divided by city size, is uniquely defined by the seven ordinal properties that characterize the M index as shown by Frankel and Volij (2009). Note that $-\log(l)$ is less appealing than $\hat{M}_T$ as a measure of segregation because the ordering induced by $\hat{M}_T$ is size invariant, while the ordering induced by -log(l) is sensitive to sample size for any given set of relative frequencies.

The value -log(l) can be seen to be a particular case of a general *KL* divergence test for the null hypothesis that *r* independent samples are drawn from an identical distribution, whose functional form is known up to a vector of parameters of dimension *k*. Kupperman (1957) showed that, under certain regularity conditions, this general *KL* divergence test is asymptotically distributed as chi-square distribution with $(r-1)k$ degrees of freedom.[11] For the statistical model set up in the previous section, it is possible to invoke earlier well-known results on the properties of the l statistic under the null of independence to show that $2T\hat{M}_T \xrightarrow{d} c^2_{(G-1)(N-1)}$ (see Theorem 1 in the Appendix).

In many practical situations, the hypothesis of indepedence or absence of segregation will be false, so that the relevant statistical properties for the index of segregation will be those under the true alternative. Salicrú *et al*. (1994) studied the asymptotic distribution of a family of estimators for which *KL* divided by sample size T is a limiting case. Using the delta method, they find square-root convergence to a normal distribution under the alternative. In Theorem 2 in the Appendix it is shown that, under assumptions A1, A2, A3, and A8, if $p_{gn} \neq p_{g\bullet}p_{\bullet n}$ for at least one $(g,n) \in GN^c$, then $T^{1/2}\left(\hat{M}_T - M\right)$ converges in distribution to a normal distribution with mean zero and positive variance.

[11] Morales *et al.* (1995) consider $\hat{M}_T$ as a particular case of a more general family of divergence measures between two consistent estimates of a discrete distribution. They find that, under the null, the chi-square distribution is an asymptotic approximation for all members of the family.

The asymptotic power of $\hat{M}_T$ can be estimated for fixed alternatives using this last result. Note, however, that the normal approximation will likely be poor if the sample is not large, and bootstrap inference may provide better approximations to the small sample distribution of $\hat{M}_T$.

III.2. Estimation and Asymptotics of Conditional Segregation

When a sample of *iid* observations of size T is available, estimation of decomposition (3) can be carried out using the principle of analogy. The following four estimators will be considered:

$$\hat{M}_T^B(\hat{g},\hat{a}) = T^{-1}\sum_{i=1}^{T}\left\{\sum_{g=1}^{G} f(g\,|\,\mathbf{x}_i;\hat{a})\log\left(\frac{f(g\,|\,\mathbf{x}_i;\hat{a})}{\hat{p}_{g\bullet}}\right)\right\},$$
$$\hat{M}_T^W(\hat{a},\hat{b}) = T^{-1}\sum_{i=1}^{T}\left\{\sum_{g=1}^{G}\sum_{n=1}^{N} f(g,n\,|\,\mathbf{x}_i,\hat{b})\log\left(\frac{f(g,n\,|\,\mathbf{x}_i;\hat{b})}{f(g\,|\,\mathbf{x}_i;\hat{a})f(n\,|\,\mathbf{x}_i;\hat{b})}\right)\right\},$$
$$\hat{M}_T^W(\hat{g},\hat{a}) = \hat{M}_T - \hat{M}_T^B(\hat{g},\hat{a}),$$
$$\hat{M}_T(\hat{g},\hat{a},\hat{b}) = \hat{M}_T^B(\hat{g},\hat{a}) + \hat{M}_T^W(\hat{a},\hat{b}),$$

where $(\hat{g},\hat{a},\hat{b})$ are estimates for (g,a,b). For the case in which all covariates $\mathbf{x}$ are discrete and partition the set of schools, the sample analogues of decomposition (4) require no functional form assumptions for the densities of the variables. In this case:

$$\hat{M}_T^B(\hat{g},\hat{a}) = \hat{M}_T^B = \sum_{d=1}^{K}\sum_{g=1}^{G}\hat{p}_{g\bullet d}\log\left(\frac{\hat{p}_{g\bullet d}}{\hat{p}_{g\bullet}\hat{p}_{\bullet\bullet d}}\right)$$
$$\hat{M}_T^W(\hat{a};\hat{b}) = \hat{M}_T^W(\hat{g};\hat{b}) = \hat{M}_T^W = \sum_{k=1}^{K}\hat{p}_{\bullet\bullet d}\sum_{n\in d}\sum_{g=1}^{G}\hat{p}_{gn|d}\log\left(\frac{\hat{p}_{gn|d}}{\hat{p}_{g\bullet|d}\hat{p}_{\bullet n|d}}\right)$$

and $\hat{M}_T = \hat{M}_T^B + \hat{M}_T^W$ always.

The remainder of this section is devoted to the properties of these estimators. Results for the case in which all covariates are discrete and for the case in which at least one covariate is not discrete are presented in III.2.1 and III.2.2, respectively.

III.2.1. Estimation and Asymptotics of District versus School Segregation

Decomposition (4) aims to answer to what extent race segregation at district level can explain a significant amount of school segregation by race. In this section, the statistical properties of the estimators for decomposition (4), $\hat{M}_T^B$ and $\hat{M}_T^W$, are studied.[12]

The term $\hat{M}_T^B$ is itself a mutual information index so that it converges in probability to the *KL* measure M^B and, by Remark 2, can be motivated as the likelihood-ratio test for the independence between race and district membership. Asymptotic distributions for $\hat{M}_T^B$ both in the presence and the absence of dependence between race and district membership can be obtained using Theorems 1 and 2 in the Appendix after a trivial change in notation.

The within-term $\hat{M}_T^W$ can also be motivated as a likelihood-ratio test. Consider testing for the independence of race and school membership within any district *d*, i.e. $H_0 : p_{gn_d|d} = p_{g\bullet|d} p_{\bullet n_d|d}$, $(g, n_d) \in GN_d^c$, $d = 1,...,D$, versus the alternative $H_1 : p_{gn_d|d} \neq p_{g\bullet|d} p_{\bullet n_d|d}$ for at least one combination (g, n_d, d). Let $l\left(\left\{\hat{p}_{gn|d}\right\},\left\{\hat{p}_{\bullet\bullet d}\right\}\right)$ be the log-likelihood evaluated at the ML estimator, and let $l\left(\left\{\hat{p}^0_{gn|d}\right\},\left\{\hat{p}^0_{\bullet\bullet d}\right\}\right)$ be the log-likelihood for the model under H_0 evaluated at the restricted ML estimator, so that $-2\log(l^W) = -2\left(l\left(\left\{\hat{p}^0_{gn|d}\right\},\left\{\hat{p}^0_{\bullet\bullet d}\right\}\right) - l\left(\left\{\hat{p}_{gn|d}\right\},\left\{\hat{p}_{\bullet\bullet d}\right\}\right)\right)$ is the log-likelihood ratio statistic and l^W is the likelihood ratio.

Remark 3: Suppose that assumptions A1 to A8 hold, and that the vector **x** includes only district code *d*. Then $\hat{M}_T^W = \dfrac{-\log(l^W)}{T}$.

Remark 3 provides an intuitive statistical interpretation for $\hat{M}_T^W$. We are not aware of any other within-groups term in a decomposition of an index of segregation so closely related to a classical statistical test. Remark 3 can be applied to any cluster of school districts, such as cities or regions, so that

[12] By interchanging the notation for groups and organizational units, the results presented here can be applied to decompositions when the set of racial groups is partitioned into supergroups.

the within terms in the resulting decompositions can be interpreted as monotone transformations of likelihood-ratio tests for the independence between race and school membership within the districts of the corresponding cluster.

The discussion of the discrete case ends with two results which characterize the asymptotic properties of $\hat{M}_T^W$. Firstly, it is shown in Theorem 3a in the Appendix that, under general conditions, if $p_{gn|d} = p_{g\bullet|d}\, p_{\bullet n|d}$ for all (g, n, d), then $T\hat{M}_T^W$ converges in distribution to a quadratic form. Intuitively, under absence of segregation, the index of segregation for each district converges by Theorem 1 in the Appendix to a chi-square distribution. Since the within term $T\hat{M}_T^W$ is a weighted average of these terms, it does not generally converge to a chi-square distribution. Secondly, if $p_{gn|d} \neq p_{g\bullet|d}\, p_{\bullet n|d}$ for at least one (g, n, d), then $T^{1/2}\left(\hat{M}_T^W - M^W\right)$ converges in distribution to a normal distribution with zero mean and positive variance (see Theorem 3b in the Appendix).

III.2.2. Conditional Segregation with Non-Discrete Covariates

Although decomposition (3) includes decomposition (4) as a special case, the former cannot be considered a true generalization of the latter. The reason is that, while in the finite-covariates situation no restrictive functional-form assumptions for the conditional densities are required to implement decomposition (3), in the presence of countable or continuous covariates identifying functional-form assumptions are implicit in parametric assumptions A4 to A7. This has two important implications. Firstly, the sum $\hat{M}_T^B(\hat{g}, \hat{a}) + \hat{M}_T^W(\hat{a}, \hat{b})$ need not be equal to $\hat{M}_T$ for small samples. Second, there is generally no monotonous relation between the likelihood-ratio test and $\hat{M}_T^B(\hat{g}, \hat{a})$ and $\hat{M}_T^W(\hat{a}, \hat{b})$.

Clearly, the fact that $\hat{M}_T^B(\hat{g}, \hat{a})$ and $\hat{M}_T^W(\hat{a}, \hat{b})$ are unrelated to likelihood-ratio tests does not imply that they cannot be interpreted as statistical tests. The asymptotic properties of both estimators are next studied under different hypotheses, therefore providing their asymptotic motivation as statistical tests. Sufficient conditions for asymptotic normality for $T^{1/2}\left(\hat{M}_T^B(\hat{g},\hat{a}) - M^B(g_0,a_0)\right)$ and $T^{1/2}\left(\hat{M}_T^W(\hat{a},\hat{b}) - M^W(a_0,b_0)\right)$ are given in the Appendix in Theorem 4 and Theorem 5, respectively. Asymptotic normality for $\hat{M}_T^W(\hat{g},\hat{a}) = \hat{M}_T - \hat{M}_T^B(\hat{g},\hat{a})$ and $\hat{M}_T(\hat{g},\hat{a},\hat{b}) = \hat{M}_T^B(\hat{g},\hat{a}) + \hat{M}_T^W(\hat{a},\hat{b})$ follows directly. Note that these results apply to any set of consistent estimators for (g,a,b). For ML estimators, the sufficient conditions will be satisfied under general regularity conditions.

IV. U.S. SCHOOL SEGREGATION

During the past decades, the U.S. has become increasingly racially and ethnically diverse, due to higher fertility and/or immigration rates among minorities, which have led to a faster population growth than that of the white population. The demographic advances of Hispanics and Asians are concentrated in certain parts of the country, while scarcely apparent in many others.[13] The main objective of this section is to illustrate the usefulness of the decompositions proposed in this paper to analyze racial school segregation patterns under the described changing environment.

IV.1. Data

We use the Common-Core of Data (CCD) compiled by the National Center for Educational Statistics (NCES). This dataset contains school enrolment records according to racial/ethnic group from all public schools in the United States. Results are reported for the school years 1989-90 (the first year for which complete enrolment data is available) and 2005-06. Schools are retrospectively assigned Core Based Statistical Area (CBSA) codes based on 2005 ZIP codes so that comparisons over time can be made, without changes in city boundary definitions affecting the results.[14] The sample is restricted to

[13] The terms *white*, *black*, *Asian* and *Native American* are used throughout this section to refer to non-Hispanic members of these racial groups. Asians include Native Hawaiian and Pacific Islanders; Native Americans include American Indians and Alaska Natives (Innuit or Aleut). The term *Hispanic* is an ethnic rather than a racial category since Hispanic persons may belong to any race. The term racial group is used throughout to refer to each of these five racial/ethnic categories.
[14] CBSAs were published by the Office of Management and Budget in 2003 and refer collectively to urban clusters of at least 10,000 people. They replace the Metropolitan Statistical Areas (MSA s) which were used during the 1950-2000 period.

open regular schools[15] located in 960 CBSA codes –referred to as "cities"– in the 50 states and the District of Columbia. This covers approximately 74% of the student population attending U.S. public schools in 2005.

There is full information for all targeted schools in 2005. For 1989, however, a number of schools in a few states failed to report the data. As a consequence, data for 1989 only includes 839 cities.[16] Unless otherwise specified, results pertain to those schools for which racial and ethnic information is available both in 1989 and in 2005. Focusing on the schools which provide information in both years probably gives a fairer comparison between the distributions observed in 1989 and in 2005, since it does not include those schools which reported in 2005 but had failed to do so in 1989. However, interpretability of the results is also potentially compromised by the fact that some new schools were created whilst others disappeared between 1989 and 2005. Nevertheless, use of all observations does not significantly change the results (available upon request), suggesting that the selection mechanisms at work are not driving the results of our analysis.

IV.2. Racial School Segregation in the U.S.: 1989 and 2005

Table 1 presents the 1989 and 2005 school enrolment by race, the overall racial compositions in the U.S. urban public schools, and the M index of overall racial school segregation. Native American, Asian, black, and Hispanic students already made up 34.8% of the total enrolment in 1989. Since growth rates in minority enrolment were larger than among whites, by 2005 minorities accounted for almost half, 48.05%, of total enrolment. Although all minority racial groups increased their share at the expense of white students, the largest increases are by far from Hispanics, who, in 2005, were already the largest minority group in U.S. public schools.

The last row of Table 1 presents the M index, which measures the expected information of the

[15] These are all operational schools, except those focused on vocational, special, or other alternative types of education.
[16] Reardon *et al.* (2000) study the public school population between 1989 and 1995 in 217 out of 323 MSAs, as defined by the Census Bureau in 1993. Frankel and Volij (2009) present results only for the 2005/2006 school year, restricting the sample to districts in CBSAs with at least two schools which serve grades K-12. Thus, the three papers study a similar phenomenon, although ours covers a larger population during a longer period.

message that transforms the set of U.S. racial shares presented in the previous panel of Table 1 to the set of schools' racial shares. In 1989, the index of segregation (multiplied by 100) was 43.92, and the index increased by 11.3% to 48.90 between 1989 and 2005.

Table 1

IV.3. District vs. School Segregation in the U.S.

Schools are organized into a set of school districts which are themselves organized into a set of cities.[17] Thus, the overall index of segregation can be decomposed into three terms. The first term results from differences in racial shares between the cities and the national racial shares, so that it can be referred to as BC (Between Cities segregation). The second term captures differences in racial shares between the cities and the educational districts, and is referred to as WC (Within Cities segregation). Finally, the last term in the decomposition captures differences in racial shares between the districts and the schools, and is referred to as WD (Within Districts segregation). Table 2 presents the decomposition of overall racial school segregation into the three components both for 1989 and 2005. The results are in line with those reported in previous empirical studies. Firstly, the BC term, closely linked to parental choices of residence at city level, contributes the most to overall school segregation. Secondly, the WD term, closely linked to the district educational authorities' decisions, is a relatively small part –around 19%– of overall segregation.

Table 2

Theorems 2 and 3 from the Appendix can be invoked to justify the use of resampling methods. Table 2 presents 5% confidence intervals based on the normal approximation. Upper and lower limits are obtained from bootstrap estimates of the variance, using 250 bootstrap samples of each individual student racial status within schools. Given the very high level of aggregation and the large sample sizes, it is hardly surprising to confirm that all terms are significantly different from zero. Looking at

[17] The data originally consist of 5,834 districts in 1989 and 7,704 districts in 2005. For the common sample there are 5,429 districts.

differences between the 1989 and 2005 results, there is supportive evidence that the increases observed in all terms are also significant, and of the same order of magnitude: close to 12% for BC, 11.5% for WC, and 10% for WD.[18]

Aggregation at national level may mask large differences in segregation at city and district level. By Remark 3, both WC and WD can be interpreted as likelihood ratios for the null of absence of segregation in any of the more than 800 cities, or the more than 5,000 districts. Clearly, this does not imply that there is segregation in all cities and all districts. A direct way to find out how many cities and districts have significant levels of segregation is to look directly at each city and each district's local index of racial segregation. By Remark 2, these indices are formal tests for the independence of racial and organizational unit status within each geographical cluster. The distribution of these local indices under the absence of segregation can be approximated using the chi-square distribution with the appropriate degrees of freedom. A naïve procedure to assess how many cities and districts present significant levels of segregation applies the test to each city and district and then counts those cities and districts for which the null cannot be rejected at a given confidence level. Using this procedure with a 1% confidence interval, segregation is found to be significant in all cities and the vast majority (99%) of districts in both years. This approach has the well-known drawback that, by design, we should expect a positive number of rejections even if the null is always true. Several corrections have been proposed in the literature (see, for example, Romano *et al*., 2008). Using the Holm correction, segregation remains significant in all cities and in most districts, although the percentage of districts for which segregation is significantly different from zero slightly decreases (98%).

A related question is whether segregation levels are very different among cities and districts. Given that the M index represents a unique ordering of clusters of the organizational units satisfying a set of desirable properties, it is useful to address this question by assessing whether the ranking of cities and the ranking of districts is significant. There are several ways to define ranking significance. For brevity,

[18] Using all schools in 1989 and 2005 results in slightly larger increases for WC: around 16%.

here we only mean whether the position in the ranking for each of the cities and each of the districts is precisely estimated. One simple way to address this issue is by bootstrapping the rankings and reporting basic bootstrap confidence limits for the ranking for each city and district. Figure 1 presents this information graphically. The *y*-axis for each plot shows both the index values and 10% bootstrap confidence limits for each city and district by year. Cities and districts with the lowest levels of segregation are ranked first so that they are represented to the left on the *x*-axis. Thus, all graphs present a positive slope by construction. Figure 1 shows that school districts with large segregation values tend to be ranked more precisely than school districts with low segregation values. The rank of those districts with the lowest levels of segregation is, in fact, very poorly estimated and its confidence intervals often range in the hundreds of positions. Regarding cities, however, the availability of large samples allows us to obtain precise estimates of the rank in most cases. Finally, a note of caution is due regarding the interpretation of Figure 1: since the ordering is specific for each year, Figure 1 does not show ranking dynamics between 1989 and 2005.

Figure 1

IV.4. Multigroup Conditional School Segregation: The Role of Income, Wages, and Teachers per Pupil.

This subsection considers to what extent the measures of WC and WD presented so far are due to the statistical association between racial group membership and socioeconomic covariates using the methodological framework developed in subsection III.2.2. We focus on two sets of controls. Firstly, it has been argued in section II.2 that, given that household income is a potential determinant of residential and school choice, it would be interesting to identify the extent to which multigroup school segregation arises from income differences across races. In addition, residential choices may potentially be affected by the composition of earnings into wage and non-wage income in the presence of credit market restrictions. Therefore, we would hope to identify the extent to which multigroup school segregation arises from race differentials in the wage to income ratio. Secondly, the impact of class size, or its inverse –the number of teachers per pupil– on academic performance and other outcomes has long been subject to debate in academic studies and political circles, where the reduction of class sizes is frequently seen as an operational way for educational authorities to effectively increase resources in schools with special needs. At the same time, parents aware of the potential positive effects of small class size on their children's educational achievements will likely make their residential and school choices dependent on how schools differ in this dimension. Thus, it would be interesting to identify the extent to which multigroup school segregation arises from class size differences across schools. Our empirical illustration tentatively addresses these two issues by merging the CCD data with aggregated measures of income and wages at county level. We specifically study the contribution to the measurements of WC and WD in 2005 of the discrepancy in the racial mix by city and by district for different values of average annual income per capita, at county level average annual wages per job at county level, and teachers per pupil at school level.[19]

Both for WC and WD, we estimate components $M^B(g,a)$ and $M^W(g,a) = M - M^B(g,a)$ using estimates of conditional densities $f(g\,|\,\mathbf{x};a)$ based on logistic regressions carried out at district and school level. In particular, for each racial group at city (district) level we assume that $f(g_i\,|\,\mathbf{x}_i;a) = \left(1+e^{-a_g'\mathbf{x}_i}\right)^{-1}$ where $\mathbf{x}_i$ includes per capita income, wages per job, and teachers per pupil at county (school) level in addition to dummy variables for city (district) to control for between city (district) segregation. Logistic regressions for each of the five racial groups are run, using as the dependent variable in each of the regressions the logistic transformation of the observed frequency of

[19] Since income includes non-wage income, income and wages are not perfectly collinear. The variable *teachers per pupil* at school level can be constructed using the information on the number of teachers and pupils reported by most schools since 2002. County codes, also available from 2002 onwards, allow us to merge the 2005 dataset with the 2004 annual per capita personal income and average wage per job by county published by the Bureau of Economic Activity of the U.S. Department of Commerce. In our county sample, the correlation between the two variables is 0.73.

students of a given race in a given district (school), and as controls the averages at district (school) level for income, wages, and teachers per pupil, in addition to the city (district) dummies. Table 3 presents a summary of the results.

Table 3

Estimated marginal effects can be interpreted as the expected change in probability (in percentage terms) associated with a one-percentage increase in each of the controls. For example, a 1% increase in per capita personal income at district level is associated with a 0.83% expected increase in the probability of a student being white, and a 0.32% expected decrease in the probability of a student being black. At district level, increases in per capita income *ceteris paribus* are associated to increases in whites and Asians and decreases in blacks and Hispanics, whilst increases in wages per job are associated to decreases in whites and increases in blacks and Hispanics. These results arguably reflect both the higher probability of black and Hispanic students having parents who have lower overall per capita income and who are more likely to be salaried workers. Increases in teachers per pupil are associated with decreases in all minority groups. With respect to whites, the point estimate of the relation is positive, although statistically not significant. At school level, increases in county per capita income are associated again with significant increases in whites and significant decreases in blacks. The signs of the estimates for the other groups are similar to those obtained for the district level regression, but the estimates are not significant. With respect to wages per job, results are again similar for whites, Hispanic and blacks, while the parameter estimates for Native American and Asians are not significant. Finally, the effect of increases in teachers per pupil is reversed for blacks at school level: a 1% increase in the teachers per pupil in a school increases the probability that a given student is black by 0.05%. Obviously, a causality interpretation should not be attached to these estimates. Nevertheless, the strong significance of these effects suggests that a significant part of WC and WD stems from the statistical association of these covariates with race. This central issue is addressed in the last panel of Table 3.

Once estimates $\hat{a}$ are obtained using the logistic regressions carried out at district and school level, the term $M^B(g, a)$ can be estimated using $\hat{M}_T^B(\hat{g}, \hat{a})$ with $\hat{g} = (\hat{p}_{1\bullet}, ..., \hat{p}_{G\bullet})'$. The term "All controls" represents $\hat{M}_T^B(\hat{g}, \hat{a})$ at city and district level (segregation at city and district level stemming from the statistical association between race membership and per capita personal income, wages per job, and teachers per pupil). Asymptotic standard errors (using the results from Theorem 4 in the Appendix) are shown in parenthesis. Results show that most (around 64%) of WC while over 20% of WD is accounted for by these three covariates. These effects are significant even in the within-districts case. Finally, to evaluate the potentially attenuating effect on segregation of teachers per pupil, the conditional segregation terms are simulated as if this control had no effect. The average effect remains the same for WD, while it decreases very slightly for WC.

V. CONCLUSIONS

The starting point of this paper is the use of the Kullback-Leibler notion of discrepancy (Kullback and Leibler, 1951) to propose a decomposition of the Mutual Information index of segregation, *M*, first introduced by Theil and Finizza (1971) to isolate segregation conditional on any vector of socioeconomic characteristics. Estimators for *M* and the terms in its decomposition are proposed, and their asymptotic properties are obtained. The usefulness of the approach is illustrated by looking at patterns of multigroup school segregation in the U.S. for the 1989-90 and 2005-06 school years. Several interesting results stem from direct application of the tools developed in the paper.

Overall multigroup school segregation, which is measured as the discrepancy between the set of U.S. racial shares to the set of schools' racial shares, is significantly positive and has significantly increased during the 15-year period. In the decomposition of overall segregation into between-cities, within-cities, and within-districts segregation, the findings are in line with previous studies: between-cities segregation, closely linked to parental choices of residence at city level, contributes the most to overall school segregation. In contrast, within-districts segregation, potentially linked to policies by the district educational authorities, represents around 19% of overall segregation. All terms in the decomposition of overall segregation are significantly different from zero, and evidence is found that all of them significantly increased during the period.

Aggregation at national level may mask large differences in segregation at city and district level. However, when segregation is studied recursively by city and district, it is found to be significant in all cities and the vast majority of districts in both years. A related question is whether the ranking of cities and the ranking of districts is significant. Using bootstrap techniques, it is found that the rank of those districts with the lowest levels of segregation is, in fact, very poorly estimated and confidence intervals often range in the hundreds of positions. Regarding the ranking of cities, however, the availability of large samples allows us to obtain precise rank estimates for most cities. Finally, we study to what extent the measures of within-cities and within-districts segregation are due to the statistical association between racial group membership and three continuous variables: annual per capita county income, wages per job at county level, and teachers per pupil at school level. Results show that around 64% and 20% of, respectively, within-cities and within-districts segregation is accounted for by these three covariates, and that the effects are strongly significant. These results illustrate why, for both explanatory and policy reasons, it is important to identify the extent to which the value of segregation arises from income and other socioeconomic characteristics. Results suggest that to reduce school segregation levels it may prove necessary to reduce income and other inequalities across races.

APPENDIX

Proposition 1: Under assumptions A1 to A7, $M = M^B(g, a) + M^W(a, b)$ where

$$M^B(g, a) = \int_{\mathbf{x} \in \Lambda} f(\mathbf{x}) \left\{ \sum_{g=1}^{G} f(g \mid \mathbf{x}; a) \log\left(\frac{f(g \mid \mathbf{x}; a)}{p_{g\bullet}} \right) \right\} d\mathbf{x}$$

$$M^{W}(a,b)=\int_{\mathbf{x}\in\Lambda} f(\mathbf{x})\left\{\sum_{g=1}^{G}\sum_{n=1}^{N} f(g,n\,|\,\mathbf{x};b)\log\left(\frac{f(g,n\,|\,\mathbf{x};b)}{f(g\,|\,\mathbf{x};a)\,f(n\,|\,\mathbf{x};b)}\right)\right\}d\mathbf{x}.$$

Proof. First note that, given assumptions A4 and A5, $\dfrac{f_1(g,\mathbf{x})}{f_2(g,\mathbf{x})}=\dfrac{f(g\,|\,\mathbf{x};a)}{p_{g\bullet}}$. and thus

$$\int f_1(g,n,\mathbf{x})\log\left(\frac{f_1(g,\mathbf{x})}{f_2(g,\mathbf{x})}\right)dm=\int_{\mathbf{x}\in\Lambda} f(\mathbf{x})\left\{\sum_{g=1}^{G} f(g\,|\,\mathbf{x};a)\log\left(\frac{f(g\,|\,\mathbf{x};a)}{p_{g\bullet}}\right)\right\}d\mathbf{x}$$
$$=M^{B}(g,a).$$

From assumptions A5 and A6, $f_1(n\,|\,g,\mathbf{x})=\dfrac{f(n,g\,|\,\mathbf{x};b)}{f(g\,|\,\mathbf{x};a)}$ and $f_1(g,n,\mathbf{x})=f(g,n\,|\,\mathbf{x};b)f(\mathbf{x})$. Given assumption A7 we directly see that:

$$\int f_1(g,n,\mathbf{x})\log\left(\frac{f_1(n\,|\,g,\mathbf{x})}{f_2(n\,|\,g,\mathbf{x})}\right)dm=\int_{\mathbf{x}\in\Lambda} f(\mathbf{x})\left\{\sum_{g=1}^{G}\sum_{n=1}^{N} f(g,n\,|\,\mathbf{x};b)\log\left(\frac{f(g,n\,|\,\mathbf{x};b)}{f(g\,|\,\mathbf{x};a)\,f(n\,|\,\mathbf{x};b)}\right)\right\}d\mathbf{x}$$
$$=M^{W}(a,b).$$

■

Proposition 2: Under assumptions A1 to A3, plim $\hat{M}_T=M$.

Proof. We first note that by direct application of Lemma 1 in Rao (1957), the sample frequencies $\hat{p}_{gn}=\dfrac{T_{gn}}{T}$ converge in probability to the actual probabilities, i.e. plim $\hat{p}_{gn}=p_{gn}$. Define the mutual information index as a function of the parameter vector $m(q)=\sum_{GN^c} p_{gn}\log\left(\dfrac{p_{gn}}{p_{g\bullet}(q)\,p_{\bullet n}(q)}\right)+\left(1-\sum_{GN^c} p_{gn}\right)\log\left(\dfrac{1-\sum_{GN^c} p_{gn}}{p_{G\bullet}(q)\,p_{\bullet N}(q)}\right)$ where $q\in\Theta$,

$$p_{g\bullet}(q)=\begin{cases}\displaystyle\sum_{n=1}^{N} p_{gn} & \text{if } g\neq G\\ \displaystyle\sum_{n=1}^{N-1} p_{Gn}+\left(1-\sum_{(j,n)\in GN^c} p_{jn}\right) & \text{if } g=G\end{cases}$$

and

$$p_{\bullet n}(q)=\begin{cases}\displaystyle\sum_{g=1}^{G} p_{gn} & \text{if } n\neq N,\\ \displaystyle\sum_{g=1}^{G-1} p_{gN}+\left(1-\sum_{(g,j)\in GN^c} p_{gj}\right) & \text{if } n=N.\end{cases}$$

Let q^{0} be the vector containing the true probabilities. Since $m(q)$ is continuous at $q=q^{0}$, by the Slutsky theorem it follows that plim $\hat{M}_T=$ plim $m(\hat{q})=m(q^{0})=M$. ■

Theorem 1: Suppose that A1, A2, A3, and A8 hold. If $p_{gn}=p_{g\bullet}p_{\bullet n}$, for all $(g,n)\in GN^c$, then:

$$2T\hat{M}_T\xrightarrow{d} c^{2}_{(G-1)(N-1)}.$$

Proof. We first note that the l statistic is also the likelihood ratio for testing H_0: $p_{g|n} = p_{g\bullet}$, $g = 1,\ldots,G-1$, $n = 1,\ldots,N$, versus the two-sided alternative H_1: $p_{g|n}$? $p_{g\bullet}$ where the quantities $T_{\bullet n}$ are assumed to be constants. Direct application of Theorem 7 in Neyman (1949) implies that $-2\log(l) \xrightarrow{d} c^2_{df}$ where $df = (G-1)N - (G-1) = (G-1)(N-1)$. The proof then follows from Proposition 2. ■

Theorem 2: Suppose that assumptions A1, A2, A3, and A8 hold. If $p_{gn} \neq p_{g\bullet}p_{\bullet n}p_{gn}$? $p_{g\bullet}p_{\bullet n}$ for at least one $(g,n) \in GN^c$, then $T^{1/2}\left(\hat{M}_T - M\right) \xrightarrow{d} N\left(0, \Delta m' \Sigma \Delta m\right)$ where

$$\Sigma = \left\{s_{(i,j)(g,n)}\right\} = \begin{cases} p_{ij}\left(1-p_{ij}\right) & \text{if } (i,j) = (g,n) \\ -p_{ij}p_{gn} & \text{if } (i,j) \neq (g,n) \end{cases}$$

$$\Delta m = \left\{\Delta m_{gn}\right\} = \log\left(\frac{p_{gn}}{p_{g\bullet}p_{\bullet n}}\right) - \log\left(\frac{p_{GN}}{p_{G\bullet}p_{\bullet N}}\right),\ \forall (g,n) \in GN^c.$$

Proof. We first note that $T^{1/2}\left(\hat{q} - q^0\right) \xrightarrow{d} N\left(0, \Sigma\right)$ (see, for example, Serfling, 1980, Theorem 2.7, p.109). To prove Theorem 2, we will use two lemmata:

Lemma 1: Suppose that $\hat{q}$ is $AN\left(q, T^{-1}\Sigma\right)$. If $m(q)$ has a non-zero partial derivative at $q = q^0$, $\Delta m \equiv \left(\left.\frac{\partial m(q)}{\partial q_1}\right|_{q=q^0}, \ldots, \left.\frac{\partial m(q)}{\partial q_{NG-1}}\right|_{q=q^0}\right)^t \neq 0_{NG-1}$, then $m(q)$ is $AN\left(m\left(q^0\right), T^{-1}\Delta m' \Sigma \Delta m\right)$.

Proof. This is a direct from Theorem 3.3A in Serfling (1980). ■

Lemma 2: For $q_i = p_{gn}$: $$\frac{\partial m(q)}{\partial q_i} = \log\left(\frac{p_{gn}}{p_{g\bullet}p_{\bullet n}}\right) - \log\left(\frac{p_{GN}}{p_{G\bullet}p_{\bullet N}}\right).$$

Proof. Define $h_{gn} = p_{gn}\log\left(\frac{p_{gn}}{p_{g\bullet}p_{\bullet n}}\right)$, where $p_{g\bullet} = \sum_{n=1}^{N} p_{gn}$, $p_{\bullet n} = \sum_{g=1}^{G} p_{gn}$, and $p_{GN} = 1 - \sum_{n=1}^{N-1}\sum_{g=1}^{G-1} p_{gn} - \sum_{g=1}^{G-1} p_{gN} - \sum_{n=1}^{N-1} p_{Gn}$. Then, for any race and school combination e and s the partial derivative of h_{es} with respect to p_{gn} is $\frac{\partial h_{es}}{\partial p_{gn}} = \left[1 + \log\left(\frac{p_{es}}{p_{e\bullet}p_{\bullet s}}\right)\right]\frac{\partial p_{es}}{\partial p_{gn}} - \left(\frac{p_{es}}{p_{e\bullet}p_{\bullet s}}\right)\frac{\partial\left(p_{e\bullet}p_{\bullet s}\right)}{\partial p_{gn}}$. The result follows after some algebraic manipulation after noticing that $\frac{\partial m}{\partial p_{gn}} = \sum_{e=1}^{G}\sum_{s=1}^{N}\frac{\partial h_{es}}{\partial p_{gn}}$. ■

In view of Lemmata 1 and 2, to prove Theorem 2 we only need to show that f p_{gn} ? $p_{g\bullet}$ $p_{\bullet n}$ for at least one $(g,n) \in GN^c$, then $\Delta m \neq 0_{NG-1}$. Since $p_{GN} = 1 - \sum_{g=1}^{G-1}\sum_{n=1}^{N-1} p_{gn} - \sum_{n=1}^{N-1} p_{Gn} - \sum_{g=1}^{G-1} p_{gN}$, if $p_{gn} = p_{g\bullet}\, p_{\bullet n}$ for all $(g,n) \in GN^c$, then $p_{GN} = p_{G\bullet}p_{\bullet N}$. Thus, if p_{gn} ? $p_{g\bullet}$ $p_{\bullet n}$ for at least one $(g,n) \in GN^c$, then there must be at least another combination (i, j) such that p_{ij} ? $p_{i\bullet}$ $p_{\bullet j}$. Assume, wlog, that p_{GN} ? $p_{G\bullet}$ $p_{\bullet N}$. We can now prove by contradiction that

$\Delta m \neq 0_{NG-1}$. Assume otherwise that $\log\left(\frac{p_{gn}}{p_{g\bullet}p_{\bullet n}}\right)=\log\left(\frac{p_{GN}}{p_{G\bullet}p_{\bullet N}}\right)\forall(g,n)\in GN^c$. Then $p_{gn}=\left(\frac{p_{GN}}{p_{G\bullet}p_{\bullet N}}\right)p_{g\bullet}p_{\bullet n}$ and summing over all $(g,n)\in GN^c$, $\sum_{GN^c}p_{gn}=\left(\frac{p_{GN}}{p_{G\bullet}p_{\bullet N}}\right)\sum_{GN^c}p_{g\bullet}p_{\bullet n}$. Given that $\sum_{GN^c}p_{g\bullet}p_{\bullet n}=1-p_{G\bullet}p_{\bullet N}$ always then $1-p_{GN}=\left(\frac{p_{GN}}{p_{G\bullet}p_{\bullet N}}\right)(1-p_{G\bullet}p_{\bullet N})$, which contradicts the initial assumption p_{GN} ? $p_{G\bullet}p_{\bullet N}$. So it must be true that if p_{gn} ? $p_{g\bullet}p_{\bullet n}$ for at least one $(g,n)\in GN^c$ then $\Delta m\neq 0_{NG-1}$. ■

Theorem 3: Suppose that assumptions A1 to A8 hold, and that the covariates vector **x** includes only district code *d*. Denote by $GN_d^c=\{(g,n_d):g=1,\ldots,G,n_d\in S_d,(g,n_d)\neq(G,N_d)\}$ the set of all race and school combinations in district *d* except combination (G,N_d). Let $q_d=(p_{11|d},\ldots,p_{G,N_d-1|d})'$ for all *d* = 1,…, *D*, and define the function $m_W(q_W)$ of parameter vector $q_W=(p_{\bullet\bullet 1},\ldots,p_{\bullet\bullet D-1},q_1{}',\ldots,q_D{}')'$ as $m_W(q_W)=\sum_{d=1}^{D-1}p_{\bullet\bullet d}m_d(q_d)+\left(1-\sum_{d=1}^{D-1}p_{\bullet\bullet d}\right)m_D(q_D)$ where

$$m_d(q_d)=\sum_{GN_d^c}p_{gn|d}\log\left(\frac{p_{gn|d}}{p_{g\bullet|d}(q_d)p_{\bullet n|d}(q_d)}\right)+\left(1-\sum_{GN_d^c}p_{gn|d}\right)\log\left(\frac{1-\sum_{GN_d^c}p_{gn|d}}{p_{G\bullet|d}(q_d)p_{\bullet N_d}(q_d)}\right)$$

for all *d* = 1,…, *D*, and $p_{g\bullet|d}(q_d)$ and $p_{\bullet n|d}(q_d)$ are defined in a similar way to $p_{g\bullet}(q)$ and $p_{\bullet n}(q)$.

(*a*) Assume further that there is at least one district *d,* such that we have that $\frac{1}{p_{gn|d}}+\frac{1}{p_{rs|d}}\neq\frac{1}{p_{g\bullet|d}}+\frac{1}{p_{\bullet n|d}}+\frac{1}{p_{r\bullet|d}}+\frac{1}{p_{\bullet s|d}}$ for at least two race and school combinations, (g,n) and (r,s).

If $p_{gn|d}=p_{g\bullet|d}p_{\bullet n|d}$ for all (g,n,d), then $T\hat{M}_T^W\xrightarrow{d}ZAZ'$ with $Z=N(0,\Sigma_W)$,

$$A=\left(\frac{1}{2}\right)\left(\left.\frac{\partial^2 m_W(q_W)}{\partial q_{Wi}\partial q_{Wj}}\right|_{q_W=q_W^0}\right)_{(GN-1)x(GN-1)},\ \text{and}\ \Sigma_W=\left\{s_{(i,j,k)(g,n,d)}\right\}=\begin{cases}p_{\bullet\bullet d}(1-p_{\bullet\bullet d}) & \text{if } (0,0,d)=(0,0,d)\\ -p_{\bullet\bullet d}p_{\bullet\bullet k} & \text{if } (0,0,d)\neq(0,0,k)\\ p_{ij|d}(1-p_{ij|d}) & \text{if } (i,j,d)=(g,n,d)\\ -p_{ij|d}p_{gn|d} & \text{if } (i,j,d)\neq(g,n,d)\end{cases}.$$

(*b*) If $p_{gn|d}\neq p_{g\bullet|d}p_{\bullet n|d}$ for at least one (g,n,d), then $T^{1/2}\left(\hat{M}_T^W-M^W\right)\xrightarrow{d}N\left(0,\Delta m^{W}{}'\Sigma_W\Delta m^W\right)$

$$\text{where } \Delta m_W=\left\{\frac{\partial m_W}{\partial q_{Wi}}\right\}=\begin{cases}m_d(q_d)-m_D(q_D) & \text{if } q_{Wi}=p_{\bullet\bullet d}.\\ \log\left(\frac{p_{gn|d}}{p_{g\bullet|d}p_{\bullet n|d}}\right)-\log\left(\frac{p_{GN_d|d}}{p_{G\bullet|d}p_{\bullet N|d}}\right) & \text{if } q_{Wi}=p_{gn|d}.\end{cases}$$

Proof. We first prove part (*a*) and then prove part (*b*). In both cases, we exploit the fact that the within term M^W can be expressed as a well-behaved function of the sample relative frequencies. Since these are the ML estimators for the actual probabilities, we have (see, for example, Serfling, 1980, pp.109) that $T^{1/2}\left(\hat{q}_W-q_W^0\right)\xrightarrow{d}N\left(0,\Sigma_W\right)$ where

q_W^0 is the vector with the actual probabilities. To prove Theorem 3*a*, we will use Lemma 2 and the following corollary from Theorem 3.3B in Serfling (1980):

Lemma 3: Suppose that $\hat{m}_T \in \mathbb{R}^K$ is $AN\left(m^0, T^{-1}\Sigma_m\right)$. If $g\left(m\right)$ is a real-valued function possessing continuous partial derivatives of second order in a neighbourhood of $m = m^0$, with the first order partial derivatives vanishing at $m = m^0$, but with the second order partial derivatives not all vanishing at $m = m^0$. Then $T\left(g(\hat{m}_T) - g(m^0)\right) \xrightarrow{d} ZAZ'$ with $Z = (Z_1, ..., Z_K) \sim N\left(0, \Sigma_m\right)$ and $A = \left(\frac{1}{2}\right)\left(\left.\frac{\partial^2 g(m)}{\partial m_i \partial m_j}\right|_{m=m^0}\right)_{KxK}$.

In view of Lemma 3, to prove Theorem 3a we only need to show that: (A) $m_W(q_W^0) = 0$; (B) $\left.\frac{\partial m_W(q_W)}{\partial q_{Wi}}\right|_{q_W = q_W^0} = 0$ for all $i = 1, ..., GN-1$; (C) $\frac{\partial^2 m_W(q_W)}{\partial q_{Wi} \partial q_{Wj}}$ are continuous in a neighbourhood of $q_W = q_W^0$; and (D) at least one second partial derivative $\frac{\partial^2 m_W(q_W)}{\partial q_{Wi} \partial q_{Wj}}$ does not vanish at $q_W = q_W^0$. Condition (A) follows immediately from the definition of $m_W(q_W)$ and the assumption that $p_{gn|d} = p_{g\bullet|d} p_{\bullet n|d}$ for all (g, n, d). To show that condition (B) holds, first note that Lemma 2 can be used to show that the partial derivative with respect to $p_{gn|d}$ is zero when $p_{gn|d} = p_{g\bullet|d} p_{\bullet n|d}$ for all (g, n, d). In addition, since $\frac{\partial m_W(q_W)}{\partial p_{\bullet\bullet d}} = m_d(q_d) - m_D(q_D)$ for all $d = 1, ..., D-1$, then these derivatives also vanish when $p_{gn|d} = p_{g\bullet|d} p_{\bullet n|d}$ since then $m_d(q_d) = m_D(q_D) = 0$. It is then straightforward to show that the second order partial derivatives $\frac{\partial^2 m_W(q_W)}{\partial q_{Wi} \partial q_{Wj}}$ are continuous in a neighbourhood of $q_W = q_W^0$. Condition (D) follows directly given that for any *g* and *n* in any district d we have that $\frac{\partial^2 m_W(q_W)}{\partial p_{gn|d}{}^2} = \left(\frac{1}{p_{gn|d}} + \frac{1}{p_{GN_d|d}}\right) - \left(\frac{1}{p_{g\bullet|d}} + \frac{1}{p_{G\bullet|d}} + \frac{1}{p_{\bullet n|d}} + \frac{1}{p_{\bullet N_d|d}}\right)$. Therefore, a suficient condition for (D) to hold is that we choose $G = r$ and $N_d = s$.

To prove Theorem 3*b*, we first note that by Theorem 3.3.A in Serfling (1980), Lemma 2, and the fact that

$$\Delta m_W = \left\{\frac{\partial m_W}{\partial q_{Wi}}\right\} = \begin{cases} m_d(q_d) - m_D(q_D) & \text{if } q_{Wi} = p_{\bullet\bullet d} \\ \log\left(\frac{p_{gn|d}}{p_{g\bullet|d} p_{\bullet n|d}}\right) - \log\left(\frac{p_{GN_d|d}}{p_{G\bullet|d} p_{\bullet N|d}}\right) & \text{if } q_{Wi} = p_{gn|d}, \end{cases}$$

the result follows if it is shown that $m_W(q_W)$ has a non-zero partial derivative at $q_W = q_W^0$. Now, using a similar argument to the argument used in the proof of Theorem 2, given that $p_{gn|d} \neq p_{g\bullet|d} p_{\bullet n|d}$ for at least one (g, n, d), then $\left.\frac{\partial m_d(q_d)}{\partial q_d}\right|_{q_d = q_d^0} \neq 0_d$. Since $\frac{\partial m_W(q_W)}{\partial p_{gn|d}} = p_{\bullet\bullet d} \frac{\partial m_d(q_d)}{\partial p_{gn|d}}$ and $p_{\bullet\bullet|d} > 0$ for all *d*, then $\left.\frac{\partial m_W(q_W)}{\partial q_W}\right|_{q_W = q_W^0} \neq \mathbf{0}_W$. ■

Theorem 4: Suppose that assumptions A1 to A5 hold and that the vector of covariates **x** includes at least one countable or continuous variable. Let (g_0, a_0) be the true parameter vectors of the data generating process and define $h_B(\mathbf{x}; g, a) = \sum_{g=1}^{G} f(g \,|\, \mathbf{x}; a) \log\left(\frac{f(g \,|\, \mathbf{x}; a)}{p_{g\bullet}} \right)$. Assume that:

(R1) $h_B(\mathbf{x}; g, a)$ is differentiable with respect to (g, a), with continuous partial derivatives which are nonvanishing at (g_0, a_0).

(R2) $\mathrm{Var}_{\mathbf{x}}\left[h_B(\mathbf{x}; g_0, a_0)\right] = s_B^2 < \infty$; $\mathrm{E}_{\mathbf{x}}\left[\left.\frac{\partial h_B}{\partial g}\right|_{g_0, a_0} \right] = m_g \in \mathbb{R}^{k_g}$, $\mathrm{E}_{\mathbf{x}}\left[\left.\frac{\partial h_B}{\partial a}\right|_{g_0, a_0} \right] = m_a \in \mathbb{R}^{k_a}$,

$\mathrm{Var}_{\mathbf{x}}\left[\left.\frac{\partial h_B}{\partial g}\right|_{g_0, a_0} \right] = \Sigma_g$, $\mathrm{Var}_{\mathbf{x}}\left[\left.\frac{\partial h_B}{\partial a}\right|_{g_0, a_0} \right] = \Sigma_a$, where Σ_g and Σ_a are positive definite matrices.

(R3) $\mathrm{plim}\ \hat{g} = g_0$ and $\mathrm{plim}\ \hat{a} = a_0$.

Then, $T^{1/2}\left(\hat{M}_T^B(\hat{g}, \hat{a}) - M^B(g_0, a_0) \right) \xrightarrow{d} N\left(0, s_B^2\right)$.

Proof. We first note that by the Lindeberg-Levy central limit theorem,

$$T^{1/2}\left(\tilde{M}_T^B(g_0, a_0) - M^B(g_0, a_0) \right) \xrightarrow{d} N\left(0, s_B^2\right)$$

where $\tilde{M}_T^B(g_0, a_0) = T^{-1} \sum_{i=1}^{T} h_B(\mathbf{x}_i; g_0, a_0)$. By the mean value theorem, we have that

$$\hat{M}_T^B(\hat{g}, \hat{a}) = \tilde{M}_T^B(g_0, a_0) + T^{-1} \sum_{i=1}^{T} \left\{ \left.\frac{\partial h_B}{\partial g}\right|_{g^*, a^*} (\hat{g} - g_0) + \left.\frac{\partial h_B}{\partial a}\right|_{g^*, a^*} (\hat{a} - a_0) \right\}$$

satisfying

$\|g^* - g_0\| \leq \|\hat{g} - g_0\|$

$\|a^* - a_0\| \leq \|\hat{a} - a_0\|$.

Therefore, to prove the theorem we only need to prove that

$$T^{-1} \sum_{i=1}^{T} \left\{ \left.\frac{\partial h_B}{\partial g}\right|_{g^*, a^*} (\hat{g} - g_0) + \left.\frac{\partial h_B}{\partial a}\right|_{g^*, a^*} (\hat{a} - a_0) \right\} = \mathrm{o}(1).$$

Given (R3), this condition is satisfied if $T^{-1} \sum_{i=1}^{T} \left.\frac{\partial h_B}{\partial g}\right|_{g^*, a^*}$ and $T^{-1} \sum_{i=1}^{T} \left.\frac{\partial h_B}{\partial a}\right|_{g^*, a^*}$ are asymtotically normal, which follows by the squeeze theorem from asymptotic normality for $T^{-1} \sum_{i=1}^{T} \left.\frac{\partial h_B}{\partial g}\right|_{\hat{g}, \hat{a}}$ and $T^{-1} \sum_{i=1}^{T} \left.\frac{\partial h_B}{\partial a}\right|_{\hat{g}, \hat{a}}$, (this is due to R3 and continuity in the partial derivatives) and asymptotic normality for $T^{-1} \sum_{i=1}^{T} \left.\frac{\partial h_B}{\partial g}\right|_{g_0, a_0}$ and $T^{-1} \sum_{i=1}^{T} \left.\frac{\partial h_B}{\partial a}\right|_{\hat{g}, \hat{a}}$ (due to the central limit theorem). ■

Theorem 5: Suppose that assumptions A1 to A7 hold, and that the vector of covariates $\mathbf{x}$ includes at least one countable or continuous variable. Let (a_0, b_0) be the true parameter vectors of the data generating process, and define $h_W(\mathbf{x};a,b)=\sum_{g=1}^{G}\sum_{n=1}^{N}f(g,n\,|\,\mathbf{x};b)\log\left(\frac{f(g,n\,|\mathbf{x}\;b)}{f(g\,|\,\mathbf{x};a)f(n\;|\mathbf{x}\;b)}\right)$. Assume that:

(R4) $h_W(\mathbf{x};a,b)$ is differentiable with respect to (a,b), with continuous partial derivatives which are non-vanishing at (a_0,b_0).

(R5) $\mathrm{Var}_{\mathbf{x}}\left[h_W(\mathbf{x};a_0,b_0)\right]=s_W^2<\infty$, $\mathrm{E}_{\mathbf{x}}\left[\left.\frac{\partial h_W}{\partial a}\right|_{a_0,b_0}\right]=m_{\mathbf{a}}\in\mathbb{R}^{k_a}$, $\mathrm{E}_{\mathbf{x}}\left[\left.\frac{\partial h_W}{\partial b}\right|_{a_0,b_0}\right]=m_b\in\mathbb{R}^{k_b}$, $\mathrm{Var}_{\mathbf{x}}\left[\left.\frac{\partial h_W}{\partial a}\right|_{a_0,b_0}\right]=\boldsymbol{\Sigma}_{\mathbf{a}}$, and $\mathrm{Var}_x\left[\left.\frac{\partial h_W}{\partial b}\right|_{a_0,b_0}\right]=\boldsymbol{\Sigma}_b$, where $\boldsymbol{\Sigma}_{\mathbf{a}}$ and $\boldsymbol{\Sigma}_b$ are positive definite matrices.

(R6) plim $\hat{a}=a_0$ and plim $\hat{b}=b_0$.

Then, $T^{1/2}\left(\hat{M}_T^W(\hat{a},\hat{b})-M^W(a_0,b_0)\right)\xrightarrow{d}N\left(0,s_W^2\right)$.

Proof. It is formally similar to the proof of Theorem 4.

REFERENCES

Aslund, O., and O.N. Skans (2009), "How to measure segregation conditional on the distribution of covariates," *Journal of Population Economics* 22(4):971-981.

Boisso, D., K. Hayes, J. Hirschberg, and J. Silber (1994), "Occupational Segregation in the Multidimensional Case: Decomposition and Test of Significance," *Journal of Econometrics* **61**: 161-171.

Carrington, W.J., and K.R. Troske (1997), "On Measuring Segregation in Samples with Small Units," *Journal of Business and Economic Statistics*, **15**: 402-409.

Clotfelter, C. (1979), "Alternative Measures of School Desegregation: A Methodological Note", *Land Economics*, **54**: 373-380.

Frankel, D and O. Volij (2009), "Measuring School Segregation," mimeo, November 2009.

Flückiger, Y. and Silber, J. (1999), *The Measurement of Segregation in the Labor Force*, Heidelberg, Physica-Verlag.

Hellerstein, J.K., and D. Neumark (2008), "Workplace Segregation in the United States: Race, Ethnicity, and Skill," *Review of Economics and Statistics* **90(3)**: 459-477.

Herranz, N., Mora, R., and Ruiz-Castillo, J. (2005), "An Algorithm to Reduce the Occupational Space in Gender Segregation Studies," *Journal of Applied Econometrics* **20**: 25-37

James, D.R. and K.E. Taeuber (1985), "Measures of Segregation," in G. Schmid and R. Weitzel (eds.), *Sex Discrimination and Equal Opportunity: The Labor Market and Employment Policy*, London, Gower Publishing Company.

Kakwani, N.C. (1994), "Segregation by Sex: Measurement and Hypothesis Testing," *Research on Economic Inequality* **5**: 1-26.

Kalter, F. (2000), "Measuring Segregation and Controlling for Independent Variables," Working Paper 19-2000, Manheimer Zentrum für Europäische Sozialforschung.

Kullback, S. (1959), *Information Theory and Statistics*, John Wiley and Sons, NY.

Kullback, S., and R.A. Leibler (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics* **22**(1):79-86.

Kupperman, M. (1957), "Further applications of information theory to multivariate analysis and statistical inference," *Dissertation*, Graduate Council of George Washington Univ.

Massey, D. and N. Denton (1988), "The Dimensions of Residential Segregation," *Social Forces* **67**: 281-315.

Mora, R. and Ruiz-Castillo, J. (2003), "Additively Decomposable Segregation Indices. The Case of Gender Segregation By Occupations and Human Capital Levels In Spain," *Journal of Economic Inequality* **1**: 147-179.

Mora, R. and Ruiz-Castillo, J. (2004), "Gender Segregation by Occupations in the Public and the Private Sectors. The Case of Spain In 1977 and 1992," *Investigaciones Económicas* **XXVIII**: 399-428.

Mora, R. and Ruiz-Castillo, J. (2009), "Entropy-based Segregation Indices," Working Paper 09-32, Economics Series 18, Universidad Carlos III.

Morales, D., L. Pardo, I. Vajda (1995), "Asymptotic divergence of estimates of discrete distributions," *Journal of Statistical Planning and Inference* **48**(3): 347-369.

Neyman, J. (1949), "Contribution to the theory of the c^2 test," in J. Neyman (ed.) *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability* , Univ. of California Press, Berkeley, 547-566.

Rao, C.R. (1957), "Maximum Likelihood Estimation for the Multinomial Distribution," *The Indian Journal of Statistics (1933-1960)* **18**:139-148.

Reardon, S., J. Yun and T. McNulty, (2000), "The Changing Structure of School Segregation: Measurement and Evidence of Multi-racial Metropolitan Area School Segregation, 1989-1999," *Demography* **37**: 351-364.
Reardon, S. and G. Firebaugh, (2002), "Measures of Multigroup Segregation," *Sociological Methodology* **32**: 33-67.
Rivkin, S.G. (1994), "Residential Segregation and School Integration," *Sociology of Education*, **67**:279-92.
Romano, J.P., A.M. Shaikh, and M. Wolf (2008), "Formalized Data Snooping Based On Generalized Error Rates," *Econometric Theory* **24**(02):404-447.
Salicrú, M., D. Morales, M.L. Menéndez and L. Pardo (1994), "On the applications of divergence type measures in testing statistical hypotheses," *Journal of Multivariate Analysis* **51**: 372-391.
Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: John Wiley & Sons.
Shorrocks, A. (1984), "Inequality Decomposition by Population Subgroups," *Econometrica* **52**: 1369-1385.
Spriggs, W.E., and R.M. Williams (1996), "Logit Decomposition Analysis of Occupational Segregation: Results for the 1970s and 1980s," *The Review of Economics and Statistics* **78**: 348-355.
Theil, H. and A.J. Finizza (1971), "A Note on the Measurement of Racial Integration of Schools by Means of Information Concepts," *Journal of Mathematical Sociology* **1**: 187-194.
Theil, H. (1972), *Statistical Decomposition Analysis*, Amsterdam: North Holland.

Table 1. Urban Public School Enrolment, Racial Mix, and School Segregation in the U.S., 1989:2005

	No. of students (millions)			Racial Shares (%)		
	1989	2005	Change (%)	1989	2005	Change (%)
Native American	0.18	0.31	70.36	0.68	0.89	0.20
Asian	1.12	2.03	82.14	4.15	5.49	1.34
Black	4.55	6.94	52.63	16.10	17.80	1.70
Hispanic	3.75	8.39	123.95	13.85	23.87	10.02
Non-white	9.59	17.68	84.26	34.78	48.05	13.27
White	16.98	18.47	8.77	65.22	51.95	-13.27
Total	**26.57**	**36.14**	**36.03**	**100.00**	**100.00**	**0.00**

	Mutual Information index of Segregation		
	1989	2005	Change (%)
M	43.92	48.90	11.33%

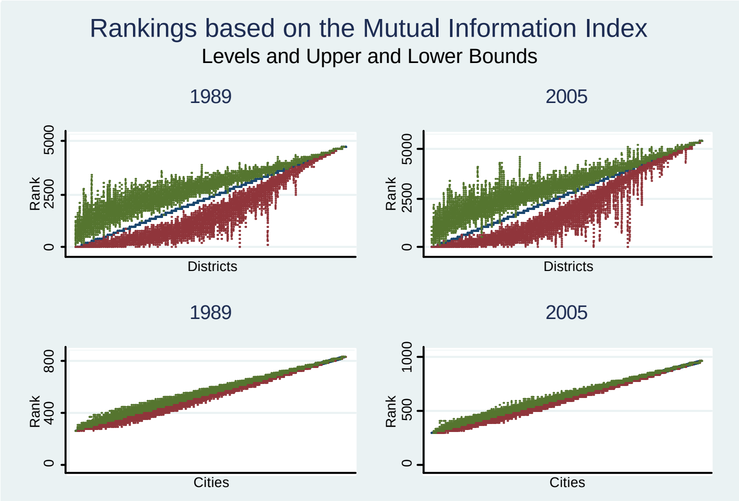
Note: Ethnic shares are the percentages of students from every race/ethnic group. The terms Native American, Asian, Black, and White refer to non-Hispanic members of these racial groups. Asian includes Native Hawaiians and Pacific Islanders; Native American includes American Indians and Alaska Natives (Innuit or Aleut). The term Hispanic is an ethnic rather than a racial category since Hispanic persons may belong to any race. Total Non-white includes all categories except White.

Table 2. Between Cities, Within Cities, and Within Districts Segregation in the U.S.

	1989			2005		
	Index	Lower Bound	Upper Bound	Index	Lower Bound	Upper Bound
Between Cities	21.04	21.02	21.06	23.50	23.48	23.52
(% over total)	47.91			48.06		
Within Cities	14.71	14.69	14.72	16.40	16.38	16.42
(% over total)	33.49			33.54		
Within Districts	8.18	8.16	8.19	9.00	8.99	9.01
(% over total)	18.62			18.40		
Total	43.92	43.90	43.95	48.90	48.87	48.92

Note: Ethnic shares are the percentages of students from every race/ethnic group. The terms Native American, Asian, Black, and White refer to non-Hispanic members of these racial groups. Asian includes Native Hawaiians and Pacific Islanders; Native American includes American Indians and Alaska Natives (Innuit or Aleut). The term Hispanic is an ethnic rather than a racial category since Hispanic persons may belong to any race. Total Non-white includes allcategories except White.

Figure 1



Notes: 5% and 95% bounds obtained from 250 bootstrap replications

Table 3. Multigroup Conditional School Segregation: The Role of Income, Wages, and Teachers per Pupil, 2005

Fixed-Effects Logistics Regressions: Marginal Effects [a]

District Level Data (7419 districts, 923 cities)

	Native American	Asian	Black	Hispanic	White
Per-capita personal income	0.00	0.04***	-0.32***	-0.09***	0.83***
Wages per job	-0.01***	-0.003	0.62***	0.14***	-1.48***
Teachers per pupil	-0.02***	-0.08***	-0.15***	-0.15***	0.10

School Level Data (53174 schools, 7419 districts)

	Native American	Asian	Black	Hispanic	White
Per-capita personal income	0.003	0.01	-0.05	-0.14***	0.18***
Wages per job	-0.004	0.03*	0.14***	0.19***	-0.35***
Teachers per pupil	-0.01***	-0.06***	0.05***	-0.01	-0.14***

The Decomposition of Within-Cities and Within-Districts Segregation [b]

	Within Cities		Within Districts	
	Index	**% over Total**	**Index**	**% over Total**
Total	16.72	100.00	9.07	100.00
All Controls	10.67 (0.003)	63.78	1.93 (0.001)	21.25
Income and Wages	10.69 (0.003)	63.92	1.93 (0.001)	21.32
Conditional Segregation	6.06	36.22	7.14	78.75

Notes:

[a] District level regressions include city fixed effects. School level regressions include district fixed effects. Marginal effects are sample averages of the estimated partial derivative of each of the controls over the probability of belonging to each of the races and can be interpreted as the expected change in probability (in percentage terms) brought about by a one-percentage increase in the control. ***, **, and * denote parameter significant at 1,5, and 10 significance level.

[b] Total-Within Cities Index measures the expected information of the message that transforms the set of city ethnic shares to the set of district ethnic shares for the regressions sample. All controls-Within Cities captures segregation stemming from the statistical association between per-capita personal income, wages per job, and teachers per pupil and race membership at district level. Income and Wages-Within Cities simulates the value of segregation stemming from the statistical association between income and wages as if teachers per pupil played no role. Conditional Segregation-Within Cities reports the difference between Total-Within Cities and All Controls-Within Cities. Total-Within District Index measures the expected information of the message that transforms the set of district ethnic shares to the set of school ethnic shares for the regressions sample. All controls-Within Districts captures segregation stemming from the statistical association between per-capita personal income, hourly wages, and teachers per pupil and race membership at school level. Income and Wages-Within Districts simulates the value of segregation stemming from the statistical association between income and wages as if teachers per pupil played no role at school level. Conditional Segregation-Within Districs reports the difference between Total-Within Districts and All Controls-Within Districts. Asymptotic Standard Errors are shown in parenthesis..