NBER WORKING PAPER SERIES

AN EQUILIBRIUM MODEL OF SORTING IN AN URBAN HOUSING MARKET

Patrick Bayer
Robert McMillan
Kim Rueben

Working Paper 10865
http://www.nber.org/papers/w10865

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2004

We would like to thank Fernando Ferreira for outstanding research assistance. Thanks also to Pedro Cerdan and Jackie Chou for help in assembling the data. We are grateful to Pat Bajari, Steve Berry, Dennis Epple, Tom Nechyba, Holger Sieg, Aloysius Siow and Chris Timmins for valuable discussions and to many others for additional suggestions, including conference participants at the AEA, ERC, IRP, NBER, PET, SITE, and SIEPR, and seminar participants at Brown, Chicago, Colorado, Columbia, Duke, Johns Hopkins, LSE, Northwestern, NYU, PPIC, Stanford, Toronto, UC Berkeley, UC Irvine, UCLA, and Yale. This research was conducted at the California Census Research Data Center; our thanks to the CCRDC, and to Ritch Milby in particular. We gratefully acknowledge financial support for this project provided by the National Science Foundation under grant SES- 0137289, the Public Policy Institute of California, and SSHRC Canada. Please send correspondence to Patrick Bayer, 37 Hillhouse Avenue, Yale University, New Haven, CT 06511, Patrick.bayer@yale.edu. The views expressed herein are those of the author(s) and not necessarily those of the National Bureau of Economic Research.

An Equilibrium Model of Sorting in an Urban Housing Market
Patrick Bayer, Robert McMillan, and Kim Rueben
NBER Working Paper No. 10865
October 2004
JEL No. H0, J7, R0, R2

ABSTRACT

This paper introduces an equilibrium framework for analyzing residential sorting, designed to take advantage of newly available restricted-access Census microdata. The framework adds an equilibrium concept to the discrete choice framework developed by McFadden (1973, 1978), permitting a more flexible characterization of preferences than has been possible in previously estimated sorting models. Using data on nearly a quarter of a million households residing in the San Francisco Bay Area in 1990, our estimates provide a precise characterization of preferences for many housing and neighborhood attributes, showing how demand for these attributes varies with a household's income, race, education, and family structure. We use the equilibrium model in combination with these estimates to explore the effects of an increase in income inequality, the findings indicating that much of the increased spending power of the rich is absorbed by higher housing prices.

Patrick Bayer
Department of Economics
Yale University
Box 208264
New Haven, CT 06520-8264
and NBER
patrick.bayer@yale.edu

Robert McMillan
University of Toronto
mcmillan@chass.utoronto.ca

Kim Rueben
Public Policy Institute of California
rueben@ppic.org

1 INTRODUCTION

Economists have long been interested in analyzing residential sorting in an urban setting. A long line of theoretical studies, including important papers by Epple, Filimon and Romer (EFR) (1984, 1993), Benabou (1993, 1996), Fernandez and Rogerson (1996, 1998), and Nechyba (1999, 2000), have developed and used models of sorting to analyze the way that interdependent individual decisions in the housing market aggregate up to determine the equilibrium structure of a metropolitan area. As these papers demonstrate, equilibrium models of residential sorting provide a coherent framework for analyzing the provision of local public goods, residential segregation, and education finance reform, proving particularly useful in tracing many complex and otherwise difficult-to-predict effects of policy.

In recent years, a new line of empirical research has sought to take these models to the data. Epple and Sieg (1999) develop an estimator for the equilibrium sorting model of EFR, providing the first unified treatment of theory and empirics in the literature. In the same vein, Sieg *et al.* (2004) use this approach to explore the general equilibrium impacts of air quality improvements in the Los Angeles Basin.[1] Concurrent with these developments, the availability of data appropriate for estimating models of residential sorting has improved dramatically with the opening of Census Research Data Centers at several locations across the United States. These centers allow researchers to access individual-level Census data at a level of geographic detail (the Census block) far smaller than has been available in the public versions of these datasets, thereby permitting researchers to characterize residential sorting much more precisely than ever before.

[1] See also Ferreyra (2003) and Walsh (2004) for empirical applications of equilibrium sorting models to education, open space policy, and urban sprawl as well as Bayer, McMillan, and Rueben (2004), Bayer Ferreira, and McMillan (2003) Timmins (2003) and Coffey (2003) for applications using the equilibrium framework of the type developed in this paper to segregation, education, global warming, and public health.

This paper introduces a new framework for analyzing residential sorting designed to take full advantage of these newly available Census microdata. In particular, we add an equilibrium concept to the empirically-flexible discrete choice framework developed by McFadden (1973, 1978) and extended in an important way by Berry *et al.* (1995). The specification permits a more flexible characterization of preferences than has been possible in the models that have been taken to the data thus far.[2] In particular, household preferences are defined over a wide range of potentially relevant housing and neighborhood attributes, including many that are endogenously determined by the sorting process itself, and preferences for each attribute are allowed to vary in a flexible way with a household's own characteristics. The resulting horizontal model of sorting permits, for example, households to have segregating racial preferences; such preferences are not possible in a vertical specification.[3] Moreover, when combined with these rich Census data, which also characterize each individual's place of work down to the block level, this flexible preference structure brings the geography of the urban housing market into the model in a natural way, as household preferences over commuting generate geographic variation in the aggregate demand for housing in neighborhoods throughout the metropolitan area, leading for example to higher property values near employment centers.

We estimate the model using data on nearly a quarter of a million households (a 1-in-7 sample) residing in the San Francisco Bay Area in 1990, developing a strategy for identifying the

[2] In terms of the previous theoretical literature, the closest antecedent to our model is that of Nechyba (1999, 2000).

[3] It is important to point out that this flexibility in our model is made possible because we abstract from issues related to local politics. As Epple, Filimon, and Romer (1993) note, incorporating local politics into models of residential sorting requires restrictions to be placed on preferences in order to guarantee the existence of an equilibrium. Accordingly, important recent papers by Epple and Sieg (1999) and Epple, Romer and Sieg (2001) estimate equilibrium models that include voting over the level of public goods, restricting households to have shared rankings over a single public goods index. We view our model as having a comparative rather than absolute advantage over the papers in that line of the literature, better suited for exploring research questions, such as those related to segregation, where a vertical restriction is inappropriate or for use in an institutional setting such as that in Californian, where Proposition 13 leaves almost no discretion over property tax rates or the level of public goods spending at the local level.

model in the presence of correlation between unobserved housing or neighborhood quality and the prices and sociodemographic composition of the neighborhood. Such correlation is likely to arise in any model of sorting whenever households observe more housing and neighborhood attributes than does the researcher, yet the resulting endogeneity problem has not been adequately addressed in the literature. The strategy that we develop builds on the boundary fixed effects approach first used by researchers seeking to deal with the correlation of school quality with unobserved neighborhood quality; we show how this approach can be applied in our setting to properly identify preferences over neighborhood sociodemographic characteristics.

The resulting estimates provide a precise characterization of preferences for many housing and neighborhood attributes as well as showing how demand for these attributes varies with a household's income, race, education, and family structure. To illustrate the power of the general equilibrium framework, we use these preference estimates along with the sorting model to explore the impact of an increase in income inequality on the housing market equilibrium. In particular, we provide estimates of the way that an increase in income for only those households in the top quartile affects stratification patterns, consumption of housing and neighborhood attributes by households at various income levels, and the implicit price of these attributes in the marketplace.

The results indicate that the increased spending of top quartile households is reflected in significantly higher housing prices, particularly for the most desirable houses and neighborhoods in the metropolitan area. As a consequence, the effects of the increased income for households in the top quartile in terms of the increased consumption of housing and neighborhood attributes are reasonably small. The consumption of housing and neighborhood attributes by households throughout the remainder of the income distribution is adversely affected by the increased

spending power of households in the top quartile, with households near but not in the top income quartile experiencing the largest adverse effects.

The rest of the paper is organized as follows: Section 2 outlines the key feature of our San Francisco Bay Area dataset, focusing on the restricted-access Census data. Section 3 describes our equilibrium model of residential sorting and Sections 4 and 5 describe the estimation procedure and the estimated preference parameters in turn. Section 6 uses the model and estimates to conduct a general equilibrium simulation designed to examine the impact of increased income inequality on the sorting equilibrium, and Section 7 concludes.

2 DATA

The analysis conducted in this paper is facilitated by access to restricted Census microdata for 1990. These restricted Census data provide the detailed individual, household, and housing variables found in the public-use version of the Census, but also include information on the location of individual residences and workplaces at a very disaggregate level. In particular, while public-use Census data specify the PUMA (a Census region with approximately 100,000 individuals) in which a household lives, the restricted data specify the Census block (a Census region with approximately 100 individuals), thereby identifying the local neighborhood that each individual inhabits and the characteristics of each neighborhood far more accurately than has been previously possible with such a large-scale data set.

For our primary analysis, we use data from six contiguous counties in the San Francisco Bay Area: Alameda, Contra Costa, Marin, San Mateo, San Francisco, and Santa Clara. We focus on this area for two main reasons: because it is reasonably self-contained, and because the area is sizeable along a number of dimensions, including over 1,100 Census tracts, and almost 39,500

Census blocks, the smallest unit of aggregation in the data. The sample consists of 242,100 households.

The Census provides a wealth of data on the individuals in the sample – race, age, educational attainment, income from various sources, household size and structure, occupation, and employment location.[4] In addition, it provides a variety of housing characteristics: whether the unit is owned or rented, the corresponding rent or owner-reported value,[5] number of rooms, number of bedrooms, type of structure, and the age of the building. We use these housing characteristics directly and in constructing neighborhood measures that characterize the stock of housing in the neighborhood surrounding each house, as well as neighborhood racial, education and income distributions based on the households within the same Census block group, a Census region containing around 10 blocks or 500 housing units. We merge additional data describing local conditions with each house record, constructing variables related to crime rates, land use, local schools, topography, and urban density. For each of these measures, a detailed description of the process by which the original data were assigned to each house is provided in a Data Appendix. The list of the principal housing and neighborhood variables used in the analysis, along with means and standard deviations, is given in the first two columns of Table 1.

[4] Throughout our analysis, we treat the household as the decision-making agent and characterize each household's race as the race of the 'householder' – typically the household's primary earner. We assign households to one of four mutually exclusive categories of race/ethnicity: Hispanic, non-Hispanic Asian, non-Hispanic Black, and non-Hispanic White.

[5] As described in the Data Appendix, we construct a single price vector for all houses, whether rented or owned. Because the implied relationship between house values and current rents depends on expectations about the growth rate of future rents in the market, we estimate a series of hedonic price regressions for each of over 40 sub-regions of the Bay Area housing market. These regressions return an estimate of the ratio of house values to rents for each of these sub-regions and we use the average of these ratios for the Bay Area, 264.1, to convert monthly rent to house value for the purposes of reporting results at the mean.

3 A MODEL OF RESIDENTIAL SORTING

We now set out an equilibrium model of a self-contained urban housing market in which households sort themselves among the set of available housing types and locations. The model consists of two key elements: the household residential location decision problem and a market-clearing condition. While it has a simple structure, the model allows households to have heterogeneous preferences defined over housing and neighborhood attributes in a very flexible way; it also allows for housing prices and neighborhood sociodemographic compositions to be determined in equilibrium.

We estimate this model using rich individual data, appealing to the notion of revealed preference - specifically that the residential location decision reveals preferences for a wide range of housing and neighborhood attributes. By examining how location decisions vary, on average, with household characteristics such as income, education, and race, one can learn how preferences for the housing and neighborhood attributes vary with these sociodemographic characteristics. Once the broad set of preference parameters in the model have been estimated, we then use the estimates and the equilibrium model to conduct a simulation designed to explore how an increase in income inequality affects the housing market equilibrium.

The Residential Location Decision. We model the residential location decision of each household as a discrete choice of a single residence from a set of house types available in the market. The utility function specification is based on the random utility model developed in McFadden (1973, 1978) and the specification of Berry, Levinsohn, and Pakes (1995), which

includes choice-specific unobservable characteristics.[6,7] Let X_h represent the observable characteristics of housing choice *h*, including characteristics of the house itself (e.g., size, age, and type), its tenure status (rented vs. owned), and the characteristics of its neighborhood (e.g., school, crime, land use, and topography). We use the notation capital letter Z_h to represent the average sociodemographic characteristics of the corresponding neighborhood, writing it separately from the other housing and neighborhood attributes to make explicit the fact that these characteristics are determined in equilibrium.[8] Let p_h denote the price of housing choice *h* and, finally, let d_h^i denote the distance from residence *h* to the primary work location of household *i*. Each household chooses its residence *h* to maximize its indirect utility function V_h^i:[9]

$$(1) \qquad \underset{(h)}{Max} \quad V_h^i = \alpha_X^i X_h + \alpha_Z^i Z_h - \alpha_p^i p_h - \alpha_d^i d_h^i + \xi_h + \varepsilon_h^i .$$

The error structure of the indirect utility is divided into a correlated component associated with each housing choice that is valued the same by all households, ξ_h, and an individual-specific term, ε_h^i. A useful interpretation of ξ_h is that it captures the unobserved quality of each housing choice, including any unobserved quality associated with its neighborhood.[10,11]

---

[6] Discrete choice applications in the urban economics literature include Anas (1982), Quigley (1985), Gabriel and Rosenthal (1989), Nechyba and Strauss (1998), Bajari and Kahn (2004). Only the latter paper includes choice-specific unobservables.

[7] Brock and Durlauf (2001, 2003) develop a number of theoretical and econometric properties for a class of discrete choice models with social interactions, focusing primarily on models where an individual's propensity to make a choice is affected by the characteristics or decisions of individuals in a reference group. The class of models studied here differs in that the utility that an individual receives in making a choice (or the propensity of an individual to make a choice) is a function of the characteristics of others making the same choice (in our context, choosing the same neighborhood). Further, we address a number of endogeneity issues that arise when some choice characteristics are observable to households but not the researcher.

[8] This component of the utility function allows for endogenous sorting on the basis of race, as in Schelling (1969, 1971), as well as other characteristics such as income and education. The assumption that utility depends on the average sociodemographic composition of the neighborhood rather than a more complicated function is made for simplicity. The identification of more general functions is certainly possible.

[9] Alternative specifications of the indirect utility function that are non-linear in housing prices could certainly be estimated, as the linear form is not essential to the model.

[10] The inclusion of a choice-specific unobservable in this specification captures the fact that many features of a given housing type or neighborhood may be unobserved by the researcher. We assume throughout the paper that ξ_h is not

Each household's valuation of choice characteristics is allowed to vary with its own characteristics, z^i, including education, income, race, employment status, and household composition. Specifically, each parameter associated with housing and neighborhood characteristics and price, α_j^i, for $j \in \{X, Z, d, p\}$, varies with a household's own characteristics according to:

$$(2) \qquad \alpha_j^i = \alpha_{0j} + \sum_{r=1}^{R} \alpha_{rj} z_r^i ,$$

with equation (2) describing household i's preference for choice characteristic *j*.

The specification of equations (1) and (2) gives rise to a horizontal model of sorting in which households have preferences defined distinctly over each choice characteristic. This contrasts with vertical models, which restrict households to have preferences over a single locational index, thereby constraining households to have the same preference ordering across locations. The additional flexibility of horizontal model is especially relevant when modeling preferences over the neighborhood racial composition, as one would certainly expect households of different races to rank neighborhoods according to their preferences very differently. The horizontal specification also captures the geography of the urban housing market very naturally, allowing households to have preferences over neighborhoods depending on the distance from their employment locations. This gives rise to variation in the aggregate demand for housing in various neighborhoods throughout the metro area, thereby increasing equilibrium housing prices in neighborhoods near employment centers.

---

sorting dependent, that is, that all relevant neighborhood amenities affected by household sorting are included as observables.

[11] Recent papers related to housing demand and neighborhood sorting including Bayer (1999), Bajari and Kahn (2004), and Ferreira (2003) find that including a choice-specific unobservable and addressing the endogeneity problem that results from its correlation with price has a significant effect on preference estimates.

Characterizing the Housing Market. As with all models in this literature, the existence of a sorting equilibrium is much easier to establish if the individual residential location decision problem is smoothed in some way. To this end, we assume that the housing market can be fully characterized by a set of housing types that is a subset of the full set of available houses, letting the supply of housing of type h be given by S_h. We also assume that each household observed in the sample represents a continuum of households with the same observable characteristics, with the distribution of idiosyncratic tastes ε^i_h mapping into a set of choice probabilities that characterize the distribution of housing choices that would result for the continuum of households with a given set of observed characteristics.[12]

Given the household's problem described in equations (1)-(2), household i chooses housing type h if the utility that it receives from this choice exceeds the utility that it receives from all other possible house choices - that is, when

$$(3) \qquad V^i_h > V^i_k \quad \Rightarrow \quad W^i_h + \varepsilon^i_h > W^i_k + \varepsilon^i_k \quad \Rightarrow \quad \varepsilon^i_h - \varepsilon^i_k > W^i_k - W^i_h \qquad \forall \quad k \neq h$$

where W^i_h includes all of the non-idiosyncratic components of the utility function V^i_h. As the inequalities in (3) imply, the probability that a household chooses any particular choice depends in general on the characteristics of the full set of possible house types. Thus the probability P^i_h that household i chooses housing type h can be written as a function of the full vectors of housing and neighborhood characteristics (both observed and unobserved) and prices {**X**, **Z**, **p**, ξ}:[13]

[12] For expositional ease and without loss of generality, let the measure of this continuum be one.

[13] For the purposes of characterizing the equilibrium properties of the model, we include an individual's employment location in z^i and the residential location in X_h.

(4) $$P_h^i = f_h(z^i, \mathbf{Z}, \mathbf{X}, \mathbf{p}, \boldsymbol{\xi})$$

as well as the household's own characteristics z^i.

Aggregating the probabilities in equation (4) over all observed households yields the predicted demand for each housing type h, D_h:

(5) $$D_h = \sum_i P_h^i \ .$$

In order for the housing market to clear, the demand for houses of type h must equal the supply of such houses and so:

(6) $$D_h = S_h, \quad \forall h \quad \Rightarrow \quad \sum_i P_h^i = S_h \quad \forall h \, .$$

Given the decentralized nature of the housing market, prices are assumed to adjust in order to clear the market. The implications of the market clearing condition defined in equation (6) for prices are very standard, with excess demand for a housing type causing price to be bid up and excess supply leading to a fall in price. In particular, given the indirect utility function defined in (1) and a fixed set of housing and neighborhood attributes, we can prove that a unique set of prices (up to scale) clears the market:

Proposition 1: If U^i_h is a decreasing, linear function of p_h for all households and $\boldsymbol{\varepsilon}$ is drawn from a continuous distribution, a unique vector of housing prices (up to a scaleable constant) solves the system of equations depicted in (6), conditional on a set of households **z** and housing and neighborhood **Z, X**, $\boldsymbol{\xi}$ characteristics. *Proof:* See Theory Appendix.

Building on Proposition 1, the following lemma is also useful for characterizing the properties of a sorting equilibrium in the housing market:

Lemma 1: If in addition to the assumptions specified in Proposition 1, U^i_h is continuous in characteristic x_h for each household *i*, the unique vector of housing prices that clears the market is continuous in **x**. *Proof:* See Theory Appendix.

In proving Proposition 1, we show that it is possible to write the solution to (6) as a contraction mapping in **p**.[14] Thus, starting from any vector **p**, an iterative process that increases the prices of houses with excess demand and decreases the prices of houses with excess supply at each iteration leads ultimately to an even spread of households across houses. Writing this market-clearing vector of prices as $\mathbf{p}^*(\mathbf{z}, \mathbf{Z}, \mathbf{X}, \xi)$, the probability that household *i* chooses house *h* can be written:

$$(7) \qquad P_h^i = f_h\left(z^i, \mathbf{Z}, \mathbf{X}, \mathbf{p}^*(\mathbf{z}, \mathbf{Z}, \mathbf{X}, \xi), \xi\right)$$

where the notation $\mathbf{p}^*(\mathbf{z}, \mathbf{Z}, \mathbf{X}, \xi)$ indicates that the set of market-clearing prices is a function of the full matrices of the household **z** and housing and neighborhood attributes $\{\mathbf{Z}, \mathbf{X}, \xi\}$.

Defining a Sorting Equilibrium. The utility function defined in equation (1) allows households to have preferences for the sociodemographic characteristics of their neighbors.[15] Using the

[14] The conditions stated in Proposition 1 provide sufficient but not necessary conditions for the existence of a unique vector of market clearing prices. For example, while reasonable, the condition that p_h enters U^i_h in a negative manner for *every* household is more stringent than is actually necessary to ensure the uniqueness result. Essentially these conditions ensure that it is possible to write the solution to the system of equations depicted in (7) as a contraction in **p**. Beyond establishing existence this is important because it makes it possible to solve quickly for market clearing prices in counterfactual simulations.

[15] Note that it is certainly possible to allow other neighborhood characteristics such as school quality and crime to depend explicitly on neighborhood sociodemographic characteristics, provided these are continuous functions of neighborhood sociodemographic characteristics. We abstract from this issue in this paper to make the exposition as straightforward as possible. In Bayer, McMillan, and Rueben (2004) we derive bounds for general equilibrium counterfactuals that account for the fact that the levels of school quality and crime in each neighborhood is affected by the re-sorting of households.

notation $h \in n$ to indicate the housing choices that belong to neighborhood *n*, the average sociodemographic composition of neighborhood *n* is given by:

$$(8) \qquad Z_n = \sum_i \sum_{h \in n} z^i \bullet P_h^i$$

Given preferences defined either directly or indirectly over the neighborhood sociodemographic composition, a *sorting equilibrium* is defined as a set of choice probabilities $\{P_h^{i*}\}$ and a vector of housing prices $\mathbf{p}^*$ such that:

i. The housing market clears according to equation (6).

ii. The set of choice probabilities $\{P_h^{i*}\}$ is a fixed point of the mapping defined in equations (7), where **Z** is formed by explicit aggregation of P_k^{j*} $\forall$ **(***j***,***k***)** according to equation (8).

This second condition ensures that, in equilibrium, each household makes its optimal location decision given the location decisions of all other households.[16]

Existence. Combining equations (7) and (8), yields the following system of equations (one for each neighborhood) that implicitly define the vector of average neighborhood sociodemographic characteristics **Z**:

$$(9) \qquad Z_n = \sum_i \sum_{h \in n} z^i \bullet P_h^i = \sum_i \sum_{h \in n} z^i \bullet f_h\left(z^i, \mathbf{Z}, \mathbf{X}, \mathbf{p}^*(\mathbf{z}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\xi}), \boldsymbol{\xi}\right) = g_n(\mathbf{z}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\xi})$$

[16] Notice that while each household actually makes a discrete location decision, we define the equilibrium in terms of the vector of choice probabilities $\{P^i_h\}$. These choice probabilities represent the distribution of location decisions made in equilibrium by the continuum of households that each household *i* represents. Note that the alternative assumption that **ε** is observed only privately along with a symmetric Bayesian Nash equilibrium concept would

Any fixed point of this mapping, $\mathbf{Z}^* = \mathbf{g(Z^*)}$ is associated with a unique vector of market clearing prices $\mathbf{p}^*$ and a unique set of choice probabilities $\{P_h^{i\,*}\}$ that together satisfy the conditions for a sorting equilibrium. In this way, finding a sorting equilibrium can be transformed into a fixed-point problem in **Z**. The existence of a sorting equilibrium then follows directly from Brouwer's fixed-point theorem:

***Proposition 2*:** If the assumptions of Proposition 1 hold and U^i_h is continuous in **Z**, a sorting equilibrium exists. *Proof:* See Theory Appendix.

Uniqueness. While it is straightforward to establish the existence of an equilibrium for the class of models described above, a unique equilibrium need not arise. Consider an extreme example in which two types of households that have strong preferences for living with neighbors of the same type must choose between two otherwise identical neighborhoods. In this case, it is easy to see that the model has multiple equilibria. In particular, two stable equilibria arise with households sorting across neighborhoods by type. When the neighborhoods are identical except for their sociodemographic composition, the matching of each household type with a particular neighborhood is not uniquely determined in equilibrium. Thus, uniqueness is not a generic property of the class of models developed above.[17]

---

allow us to define the equilibrium in terms of discrete location decisions rather than working with the choice probabilities. Existence would continue to hold under this interpretation concerning **ε**.

[17] This extreme example does give an unduly pessimistic impression of the likelihood that multiple equilibria arise in this model. Extending the simple example just described, imagine that households of one type have significantly more income than households of the other type, that the quality of one of the neighborhoods is significantly better than that of the other neighborhood in some fixed way, and that households have preferences for neighborhood quality. In this case, while strong preferences to segregate certainly ensure that households again sort across neighborhoods by type, the matching of household type and neighborhood is made much clearer by the marked differences in income and neighborhood quality. In general, a unique equilibrium will arise when the meaningful variation in the exogenous attributes of households, neighborhoods, and houses $\{Z^i, X_h, \xi_h\}$ is sufficiently rich relative to the role that preferences. See Bayer and Timmins (2003) for a formal analysis of this issue.

4 ESTIMATION

Estimation of the model follows a two-stage procedure closely related to that developed in Berry, Levinsohn, and Pakes (1995). It is helpful in describing the estimation procedure to first introduce some notation. In particular, we rewrite the indirect utility function as:

$$(10) \qquad V_h^i = \delta_h + \lambda_h^i + \varepsilon_h^i$$

where

$$(11) \qquad \delta_h = \alpha_{0X} X_h + \alpha_{0Z} Z_h - \alpha_{0p} p_h + \xi_h$$

and

$$(12) \qquad \lambda_h^i = \left(\sum_{k=1}^{K} \alpha_{kX} z_k^i \right) X_h + \left(\sum_{k=1}^{K} \alpha_{kZ} z_k^i \right) Z_h - \left(\sum_{k=1}^{K} \alpha_{kp} z_k^i \right) p_h - \left(\sum_{k=1}^{K} \alpha_{kd} z_k^i \right) d_h .$$

In equation (11), δ_h captures the portion of utility provided by housing type *h* that is common to all households, and in (12), *k* indexes household characteristics. When the household characteristics included in the model are constructed to have mean zero, δ_h is the *mean indirect utility* provided by housing choice *h*. The unobservable component of δ_h, ξ_h, captures the portion of unobserved preferences for housing choice *h* that is correlated across households, while ε_h^i represents unobserved preferences over and above this shared component.

The estimator is a two-stage procedure. The first stage selects the heterogeneous parameters λ_h and mean indirect utilities δ_h that maximize the probability that the model correctly predicts each individual's location decision conditional on the full set of observed housing and neighborhood attributes, including those endogenously determined. Formally, the validity of this first stage requires two assumptions: that the observed location decisions are individually optimal, given the collective choices made by other households and the vector of

market-clearing prices, *and* that households are sufficiently small such that they do not interact strategically with respect to particular draws on ε. This latter assumption ensures that households can each effectively integrate out the idiosyncratic preferences of all others when making their own location decisions and so that no household's particular idiosyncratic preferences affect the equilibrium. Thus the vector of idiosyncratic preferences $\boldsymbol{\varepsilon}$ is uncorrelated with the prices and neighborhood sociodemographic characteristics that arise in *any* equilibrium.

In essence, the first-stage of the estimation procedure is equivalent to a Maximum Likelihood procedure that treats housing prices and neighborhood sociodemographic characteristics as exogenous from the individual's point-of-view. Importantly, the assumption that prices and neighborhood sociodemographic characteristics are uncorrelated with the vector of idiosyncratic preferences $\boldsymbol{\varepsilon}$ does not imply that they are uncorrelated with the full error term, as we explicitly allow for a portion of unobserved preferences, ξ, that is correlated with price and endogenous neighborhood characteristics in equilibrium. This correlation is addressed in the second stage of the estimation procedure, in which the vector $\boldsymbol{\delta}$ estimated in the first stage is decomposed into components.

Operationally, for any combination of the heterogeneous parameters in λ and mean indirect utilities, δ_h, the model predicts the probability that each household i chooses house type h. We assume that ε_h^i is drawn from the extreme value distribution, in which case this probability can be written:

$$(13) \qquad P_h^i = \mathbf{exp}(\delta_h + \hat{\lambda}_h^i) \Big/ \sum_k \mathbf{exp}(\delta_k + \hat{\lambda}_k^i)$$

Maximizing the probability that each household makes its correct housing choice gives rise to the following quasi-log-likelihood function:

$$\tilde{\ell} = \sum_i \sum_h I_h^i \ln(P_h^i) \tag{14}$$

where I^i_h is an indicator variable that equals 1 if household *i* chooses house type *h* in the data and 0 otherwise. The first stage of the estimation procedure consists of searching over the parameters in λ and the vector of mean indirect utilities to maximize $\tilde{\ell}$. Notice that the quasi-likelihood function developed here is based solely on the notion that each household's residential location is optimal given the set of observed prices and the location decisions of other households.

The Mechanics of the First Stage of the Estimation. Intuitively, it is easy to see how this first stage of the estimation procedure ties down the heterogeneous parameters – those involving an interaction of household characteristics with housing and neighborhood characteristics. If more educated households are more likely to choose houses near better schools in the data for instance, a positive interaction of education and school quality will allow the model to fit the data better than a negative interaction would. What is less intuitive is the way the vector of mean indirect utilities is determined. To better understand the mechanics of the first stage of the estimation, it is helpful to write the first-order conditions related to δ_h:

$$\partial\tilde{\ell}/\partial\delta_h = \sum_{i\in h} \partial \ln(P_h^i)/\partial\delta_h + \sum_{i\notin h} \partial \ln(P_h^i)/\partial\delta_h = \sum_{i\in h}\left(1 - P_h^i\right) + \sum_{i\notin h}\left(-P_h^i\right) = S_h - \sum_i \left(P_h^i\right) = 0 \tag{15}$$

It is apparent that the quasi-likelihood function is maximized at the vector δ that forces the sum of the probabilities that each observed individual chooses each house type to equal the total supply of such houses: $\sum_i \left(P_h^i\right) = S_h \; \forall \, h$. That this condition must hold for all house types results from a fundamental trade-off in $\tilde{\ell}$. In particular, an increase in any δ_h raises the probability that each household in the sample chooses house type *h*. While this increases the probability that the model correctly predicts the choice of the households that actually reside in houses of type *h*, it decreases the probability that all of the other households in the sample make the correct choice. Thus the first stage of the estimation procedure consists of choosing the interaction parameters that best match each individual with their chosen house, while ensuring that total predicted demand equals supply for each house type.

For any set of interaction parameters (those in λ), a contraction mapping can be used to calculate the vector δ that solves the set of first order conditions: $\sum_i \left(P_h^i\right) = S_h \; \forall \; h$. For our application, the contraction mapping is simply:

$$(16) \qquad \delta_h^{t+1} = \delta_h^t - \ln\left(\sum_i \hat{P}_h^i \Big/ S_h \right)$$

where *t* indexes the iterations of the contraction mapping. Using this contraction mapping, it is possible to solve quickly for an estimate of the full vector $\hat{\delta}$ even when it contains a large number of elements, thereby dramatically reducing the computational burden in the first stage of the estimation procedure.[18]

[18] It is worth emphasizing that a separate vector δ is calculated for each set of interaction parameters – and at the optimum, this procedure returns the quasi-ML estimates of the interaction parameters and the vector of mean indirect utilities δ.

Notice that while we have not explicitly enforced the market clearing conditions derived above, the conditions that result from maximizing the quasi-likelihood function with respect to δ are identical to the market-clearing conditions shown in equation (6). Thus, there is a clear duality between the equilibrating role of prices in our characterization of equilibrium in the housing market and the way that the vector of mean indirect utilities is determined as a result of maximizing the likelihood that each household chooses its appropriate house conditional on prices and housing and neighborhood attributes.

The Second Stage. Having estimated the vector of mean indirect utilities in the first stage of the estimation procedure, the second stage involves decomposing δ into observable and unobservable components according to equation (11).[19] Because households sort across locations based in part on the portion of housing and neighborhood quality unobserved by the researcher, housing prices and neighborhood sociodemographic characteristics are almost certainly correlated with ξ_h and consequently the corresponding endogeneity problems must be confronted.

To deal with the correlation of price and unobserved housing/neighborhood quality, ξ_h, we instrument for price. The particular instrument that we develop takes advantage of an inherent feature of housing markets: that the demand for a house in a particular neighborhood is affected not only by the features of the neighborhood itself but also by the availability of alternative houses and neighborhoods in the wider region. For example, neighborhoods that possess certain amenities that are unique or difficult to replicate will command higher prices in equilibrium, partly because of this scarcity. The exogenous attributes of houses and

neighborhoods at a reasonable distance from a particular neighborhood serve as suitable instruments for price, as the attributes of these more distant neighborhoods affect equilibrium prices but not the utility derived from living in the neighborhood.[20]

In practice, the precision of the estimation is improved significantly when the logic of this IV strategy is used to construct a single variable that approximates the optimal instrument. In particular, we construct an instrument by solving for the vector of prices that would clear the market when only exogenous features of houses and neighborhoods are included in the utility function. This instrument captures the portion of housing price variation attributable to the distribution of the exogenous features of houses and neighborhoods throughout the region, summarizing this information in a single variable.

A couple of additional practical items are worth describing. First, the construction of the instrument requires an initial conjecture as to the parameters associated with exogenous housing and neighborhood attributes. We obtain such an initial conjecture for the parameters of the mean indirect utility equation by making a reasonable guess as to the price coefficient and then estimating equation (11) via OLS, bringing the price term to the left hand side of the equation. Using the resulting coefficients on $\boldsymbol{X}$ from this regression along with those obtained in the first stage, we then calculate the vector of housing prices that clears the market, $\hat{p}^{*}(X_h, Z^i)$, setting ξ_h=0 for all *h*, and including only *exogenous* choice characteristics in the model.[21] In the results

[19] Notice that the set of observed residential choices provides no information that distinguishes the components of δ. That is, regardless of the way δ is broken into components, the effect on choice probabilities is the same.

[20] Put another way, for most individuals, the relevant extent of the housing market is much larger when they are searching for a house (they might live, for example, to the north, south, east, or west of their job location) than when they actually choose a residence, in which case the characteristics of houses on the opposite side of town likely have only a minimal direct impact on utility. It is the fact that a much broader set of houses is in play during the search process that implies that the characteristics of the housing stock on the other side of town will influence equilibrium prices.

[21] To obtain the final estimates reported in the paper, we repeat this procedure using the estimated parameters from the initial estimation to construct a new price instrument for the next iteration. While using such an iterative process is not necessary to ensure consistency, in practice it ensures that the final estimates are not sensitive to our initial

reported below, we include a full set of controls for the characteristics of a house and its neighborhood as well as five variables that describe land use[22] and six variables that describe the housing stock[23] in each of the 1, 2, 3, 4, and 5 mile rings around the house. In this way, the additional information embedded in our instrument derives from the exogenous features of the housing stock and land use in a region beyond five miles from the house in question.[24]

The Endogeneity of Neighborhood Sociodemographics. A second identification issue concerns the correlation of neighborhood sociodemographic characteristics $\boldsymbol{Z}$ with unobserved housing and neighborhood quality, ξ_h. To properly estimate preferences in the face of this endogeneity problem, we adapt a technique previously developed by Black (1999) when estimating preferences for school quality. Black's strategy makes use of a sample of houses near school attendance zone boundaries, estimating a hedonic price regression that includes boundary fixed effects. Intuitively, the idea is to compare houses in the same local neighborhood but on opposite sides of the boundary, exploiting the discontinuity in the right to attend a given school.

There are, however, good reasons to think that households will sort with respect to such boundaries. Thus, while boundary fixed effects are likely to do a good job of controlling for differences in unobserved fixed factors, neighborhood sociodemographics are likely to vary discontinuously at the boundary. In this way, the use of boundary fixed effects isolates variation

---

conjecture of the coefficient on price. For this reason, we believe that this iterative procedure is likely to be more efficient than applying the procedure once, but we do not have a proof of this.

[22] That is: percent industrial, percent commercial, percent residential, percent open space, and percent other.

[23] The housing stock variables are: percent owner-occupied single family homes with 7 rooms or more; percent owner-occupied single family homes with less than 7 rooms; percent renter-occupied single family homes; percent renter-occupied units in large apartment buildings; percent of units in small apartment buildings; percent other.

[24] In first-stage price regressions, this instrument, which is derived entirely from the exogenous characteristics of the alternatives and the distribution of household characteristics in the population, adds significantly to the predictive power of these regressions. In each specification, the optimal price instrument is strongly predictive of price, over and above the set of variables included directly in $\boldsymbol{X}$, increasing the R^2 of each regression by approximately 4 percentage points.

in both school quality and neighborhood sociodemographics in a small region in which unobserved fixed features, (e.g., access to the transportation network), likely vary only slightly, thereby providing an appealing way to account for the correlation of both school quality *and* neighborhood sociodemographics with unobservable neighborhood quality

We incorporate school district boundary fixed effects when estimating equation (11). In particular, we create a series of indicator variables for each Census block that equal one if the block is within a given distance of each unique school district boundary in the metropolitan area (e.g., Palo Alto-Menlo Park).[25] To show the variation in school quality and neighborhood sociodemographics at school district boundaries, Table 1 displays descriptive statistics for various samples related to the boundaries. The first two columns report means and standard deviations for the full sample while the third column reports means for the sample of houses within 0.25 miles of a school district boundary. Comparing the first column to the third column of the table, it is immediately obvious that the houses near school district boundaries are not fully representative of those in the Bay Area as a whole. To address this problem, we create sample weights for the houses near the boundary.[26] Column 7 of Table 1 shows the resulting weighted means, indicating that using these weights makes the sample near the boundary much more representative of the full sample.

[25] A number of empirical issues arise in incorporating school district boundary fixed effects into our analysis. A central feature of local governance in California helps to eliminate some of the problems that naturally arise with the use of school district boundaries, as Proposition 13 ensures that the vast majority of school districts within California are subject to a uniform effective property tax rate of one percent. Concerning the width of the boundaries, we experimented with a variety of distances and report the results for 0.25 miles, as these were more precise due to the larger sample size.

[26] The following procedure is used: we first regress a dummy variable indicating whether a house is in a boundary region on the vector of housing and neighborhood attributes using a logistic regression. Fitted values from this regression provide an estimate of the likelihood that a house is in the boundary region given its attributes. We use the inverse of this fitted value as a sample weight in subsequent regression analysis conducted on the sample of houses near the boundary.

The fourth and fifth columns report means for houses within 0.25 miles of a boundary, comparing houses on the high versus low average test score side of the each boundary; the sixth column reports t-tests for the difference in means. Comparing these differences reveals that houses on the high side cost $53 more per month and are assigned to schools with test scores that are 43-point higher on average.[27] Moreover, houses on the high quality side of the boundary are much more likely to be inhabited by white households and households with more education and income. These types of across-boundary differences in sociodemographic composition are what one would expect if households sort on the basis of preferences for school quality. While far less significant, other housing characteristics do vary across the boundaries as well. Consequently, we expect the use of boundary fixed effects to control for much but not all of the variation in unobserved housing and neighborhood quality, thereby giving rise to better estimates of preferences for neighborhood sociodemographics and school quality.[28]

Characterizing the Housing Market – A Practical Issue. A final practical issue for estimation concerns the way the choices that characterize the housing market should be defined. This modeling decision essentially corresponds to an assumption regarding the way demand for particular houses in the market is determined. The trade-offs implicit in the required assumption can be seen using a simple example: Consider a city neighborhood with two types of housing structures, one of which is more prevalent than the other, with all houses in the neighborhood selling for the same price. To simplify this discussion, further assume that households have

[27] As described in the Data Appendix, we construct a single price vector for all houses, whether rented or owned.

[28] In terms of the estimates related to neighborhood sociodemographic characteristics, the key point about using school district boundary fixed effects rather than Census tract fixed effects is that in the boundary case we have a clear sense of what fundamentally leads to the sorting of households across neighborhoods within the region upon which the fixed effect is based. Because we control directly for that cause of the sorting - schooling in this case - we are less concerned that the variation in sorting is related to variation in unobservables within the region upon which the fixed effect is based.

identical tastes. In this case, if we characterized the choice set as the two types of structure, we would infer that the more prevalent structure provided higher mean direct utility; this is necessary to explain why more households choose that structure given equal prices. If, on the other hand, we characterized the housing market by randomly drawing a subset of the houses in the neighborhood, we would infer that all of the houses in the neighborhood offered the same utility. We do not see any strong *a priori* for making one of these choices versus the other. Moreover, given that any definition of 'type' would be based only on the limited characteristics observed in the data, we adopt the second option described above, simply characterizing housing types as the 1-in-7 random sample of the houses observed in our Census dataset. This characterization also facilitates comparisons with the hedonic price regression literature; with this characterization of the choice set, a hedonic price regression corresponds to estimating mean preferences under the assumption of no heterogeneity in household tastes.[29]

Asymptotic Properties of the Estimator. As described in McFadden (1978), an attractive aspect of the underlying IIA property for each individual is that we can estimate the model using only a sample of the alternatives not selected by the individual. This permits estimation despite having many alternatives – i.e., many distinct house types. More generally, our problem fits within a class of models for which the asymptotic distribution theory has been developed. In this sub-section, we summarize the requirements necessary for the consistency and asymptotic normality of our estimates and provide some intuition for these conditions.

In general, there are three dimensions in which our sample can grow large: H (number of housing types), N (number of individuals in the sample), or C (number of non-chosen

[29] Nothing theoretically prevents estimation of the model under an alternative assumption concerning housing choices. A comparison with corresponding hedonic price regressions is shown in Table 2 and discussed in an

alternatives drawn for each individual). For any set of distinct housing alternatives of size *H* and any random sampling of these alternatives of size *C*, the consistency and asymptotic normality of the first-stage estimates (δ, θ_λ) follows directly as long as *N* grows large. This is the central result of McFadden (1978), justifying the use of a random sample of the full census of alternatives. Intuitively, even if each household is assigned only one randomly drawn alternative in addition to its own choice, the number of times that each house type is sampled (the dimension in which the choice-specific constants are identified) grows as a fixed fraction of *N*.

If the true vector δ were used in the second stage of the estimation procedure, the consistency and asymptotic normality of the second-stage estimates θ_δ would follow as long as $H \to \infty$.[30] In practice, ensuring the consistency and asymptotic normality of the second-stage estimates is complicated by the fact the vector δ is estimated rather than known. Berry, Linton, and Pakes (2002) develop the asymptotic distribution theory for the second stage estimates θ_δ for a broad class of models that contains our model as a special case, and consequently we employ their results. In particular, the consistency of the second-stage estimates follows as long as $H \to \infty$ and *N* grows fast enough relative to *H* such that $H \log H/N$ goes to zero, while asymptotic normality at rate $\sqrt{H}$ follows as long as H^2/N is bounded. Intuitively, these conditions ensure that the noise in the estimate of δ becomes inconsequential asymptotically and thus that the asymptotic distribution of θ_δ is dominated by the randomness in ξ as it would be if δ were known.

Given that the consistency and asymptotic normality of the second stage estimates requires the number of individuals in the sample to go to infinity at a faster rate than the number

---

appendix.

[30] This condition requires certain regularity conditions. See Berry, Linton, and Pakes (2002) for details.

of distinct housing units, it is important to be clear about the implications of the way that we characterize the housing market in the paper. In particular, we characterize the set of available housing types using the 1-in-7 random sample of the housing units in the metropolitan area observed in our Census dataset. Superficially, this characterization seems to imply that the number of housing types is as great as the number of households in the sample, which appears at odds with the requirements for the establishing the key asymptotic properties of our model.

It is important to note, however, the housing market may be characterized by a much smaller sample of houses, with each 'true' house type showing up many times in our large sample. Consider, for example, using a large choice set of 250,000 housing units, when the market could be fully characterized by 25,000 'true' house types, with each 'true' house type showing up an average of 10 times in the larger choice set. On the one hand, the 250,000 observations could be used to calculate the market share of each of the 25,000 'true' house types, with market shares averaging 1/25,000 and the second stage δ regressions based on 25,000 observations. On the other hand, separate market shares equal to 1/250,000 could be attributed to each house observed in the larger sample and the second stage regression based on the larger sample of 250,000. These regressions would return exactly the same estimates, as the former regression is a direct aggregation of the latter. What is important from the point-of-view of the asymptotic properties of the model is not that the number of individuals increases faster than then number of housing choices used in the analysis, but rather that the number of individuals increases fast enough relative to the number of truly distinct housing types in the market. That the number of distinct housing types in the market grows at a rate slower than the number of households seems plausible.

5 PARAMETER ESTIMATES

Estimation of the full model proceeds in two stages, as noted, the first stage recovering interaction parameters and vector of mean indirect utilities, the second stage returning the components of mean indirect utility. The first stage of the estimation procedure returns 178 parameters on terms that interact individual and housing/neighborhood characteristics, permitting great flexibility in preferences across different types of households. In particular, the model includes the following household characteristics: household income from non-capital sources, household income from capital sources (a proxy for wealth), race, education, work status, age, the presence of children, and interactions of household income and race. These household characteristics are interacted with many housing and neighborhood attributes including house price, owner-occupancy status,[31] number of rooms, the age of the structure, average test score, elevation, population density, crime and eight variables characterizing the neighborhood sociodemographic composition: the fraction of households of each race, the fraction of households college educated, average neighborhood income, and neighborhood income interacted with race. The model also captures the spatial aspect of the housing market by allowing households to have preferences over commuting distance.[32]

Normalized estimates of the full set of parameters estimated in the first stage of the estimation procedure are reported in Appendix Table 1. To make the discussion of these estimates more transparent, we transform the estimates so that they can be described in terms of

[31] We treat ownership status as a fixed feature of a housing unit in the analysis. Thus, whether a household rents or owns is endogenously determined within the model by its house choice. In the model, we allow households to have heterogeneous preferences for home-ownership (a positive interaction between household wealth and ownership, for example, implying that wealthier households are more likely to own their housing unit, as we find below). A single price index is used for owner- and renter-occupied units - see the Data Appendix for details.

[32] We treat a household's primary work location as exogenous, calculating the distance from this location to the location of the neighborhood in question. MWTP estimates for other housing and neighborhood attributes based on a specification without commuting distance are qualitatively similar except for variables that are strongly correlated with employment access such as population density.

marginal willingness-to-pay measures (MWTP), reporting these estimates in Tables 2 and 3. The first two columns of Table 2 report measures of the mean MWTP for housing and neighborhood attributes; these estimates are based on a weighted sample of houses[33] within 0.25 miles of school district boundaries, with and without including fixed effects, respectively. Comparing the coefficients on the neighborhood sociodemographic characteristics with and without the inclusion of boundary fixed effects (columns 1 and 2) yields the pattern of results one would expect if boundary fixed effects control for fixed aspects of unobserved neighborhood quality that are correlated with neighborhood sociodemographic characteristics in the expected way. In particular, controlling for fixed effects increases the coefficient on percent black (reported at the mean average neighborhood income) from -$285 to -$234; on percent Hispanic from -$37 to $104; and on percent Asian from -$70 to $150. Doing so also reduces the coefficient on the percent of households with a college degree from $186 to $165 and the coefficient on average neighborhood income (/$10,000) from $89 to $85 per month. In this way, the use of boundary fixed effects appears to be effective in controlling for fixed aspects of unobserved neighborhood quality that are correlated with neighborhood sociodemographics, and thus provides an attractive way of estimating preferences for neighborhood sociodemographic characteristics in the presence of this important endogeneity problem.[34]

Table 3 reports the estimates of the heterogeneity in MWTP for housing and neighborhood characteristics. For ease of comparison, the first column of Table 3 reports the estimated mean MWTP for the changes in housing or neighborhood attributes described in the row headings. The remaining columns report the difference in MWTP associated with the

[33] The procedure for constructing sample weights designed to make the boundary sample as representative of the full sample as possible is described in Section 4 above. The estimates reported for the boundary sample without boundary fixed effects are qualitatively similar to those for the full sample.

comparison of household characteristics shown in the column heading. So, for example, the first entry of the table implies that, on average, households are willing-to-pay $109 more per month on the margin for an additional room, while the second entry in the first row implies that households with children are willing to pay an average of $31 per month more for a room than households without children.

In almost every instance, the parameter estimates reported in Table 3 seem to have reasonable signs and magnitudes. Focusing specifically on some of the key factors driving the location decision, the results imply that households are willing to pay an average of $50 per month to be an additional mile closer to work or about a dollar per additional mile of actual commuting travel.[35] Households with children are willing to pay more for school quality and for an extra room. A number of household characteristics in the model may proxy to some extent for lifetime wealth. Demand for larger homes, owner-occupancy (which may proxy in part for unobserved house quality), and additional rooms in a house is an increasing function of a household's income from non-capital sources, its income from capital sources, whether a household is working, and a household's educational attainment.

Turning to the estimated preferences for neighborhood attributes, an interesting distinction arises between tastes for more educated versus higher income neighbors. In particular, the estimates imply a high mean taste for neighbors with more income, but little heterogeneity in taste around this mean. Having controlled for income, however, the estimates reveal preferences for segregation on the basis of educational attainment. In particular, college-educated households are willing to pay a sizeable premium to live with other college-educated

[34] The analogous hedonic price regressions reported in the remaining columns of Table 2 provides further support for the plausibility of this assertion, as discussed in an Appendix.

households, while non-college-educated households would slightly prefer, on average, to live with others who also do not have a college degree.

In examining the estimated heterogeneity in MWTP for neighborhood racial composition, it is important to point out that the parameters corresponding to the interactions between household and neighborhood race in fact combine a number of potential explanations for racial sorting that are indistinguishable in the data. In particular, the estimated interactions combine the effects of (i) discrimination in the housing market (e.g., centralized discrimination against recent immigrants from China), (ii) direct preferences for the race of one's neighbors (e.g., preferences on the part of a recent immigrant from China to live with other Chinese immigrants), and (iii) preferences for race-specific portions of unobserved neighborhood quality (e.g., preferences for Chinese groceries which are located in neighborhoods with a high fraction of Chinese residents). If one thinks of discrimination as an expression of the racial preferences of the discriminating group concerning the group discriminated against, our model essentially mis-assigns these preferences to the group discriminated against. In this way, the estimated difference in MWTP for black versus white neighbors combines the difference that results from decentralized preferences acted upon in each individual's own location decision as well as any centralized discrimination that causes black households to appear as if they prefer black versus white neighborhoods more strongly than they actually do. Consequently, the estimates reported in Table 3 are informative about the overall importance of role of racial sorting in the housing market, but, importantly, do not distinguish preferences *per se*.

The estimated heterogeneity parameters related to race reveal strong segregating racial interactions, with, interpreted literally as preference, households of each race preferring to live

[35] Note that the estimate of each individual's disutility from commuting naturally gives rise to declining rent gradients moving away from employment centers. In this way, the model organically captures any number of

near others of the same race. In reading the numbers associated with racial interactions, it is important to keep in mind that these numbers represent the difference in the amount households of the race shown in the column heading would be willing to pay for the corresponding change in neighborhood racial composition compared with white households. So, for example, the $86 per month that characterizes the difference between the MWTP of black versus white households for a 10 percentage point increase in the fraction of black versus white neighbors reflects the sum of what a black household is willing to pay for this increase and what white households would be willing to pay for the opposite change. The parameter estimates also reveal strong segregating preferences for Hispanic and Asian households and that Asian, black, and Hispanic households are more willing to live with minority households of other races than white households are. Finally, the estimates reveal that the strength of these segregating racial interactions does not decline significantly with income. That is, high-income households of each race exhibit a remarkably similar MWTP pattern with respect to the race of their neighbors.

6 GENERAL EQUILIBRIUM SIMULATIONS

We now use the estimated parameters to conduct a general equilibrium simulation designed to examine the impact of an increase in income inequality on the housing market equilibrium. In particular, we calculate the new equilibrium that arises following a 10 percent increase in the income of households in the top income quartile, characterizing the way this change affects a number of aspects of the housing market equilibrium.

The basic structure of solving for a new equilibrium consists of a loop within a loop. The outer loop calculates the sociodemographic composition of each neighborhood, given a set of prices and an initial sociodemographic composition of each neighborhood. The inner loop

---

employment centers within the metropolitan region.

calculates the unique set of prices that clears the housing market, given an initial sociodemographic composition for each neighborhood. Thus for any change in the primitives of the model, we first calculate a new set of prices that clears the market. Using these new prices and the initial sociodemographic composition of each neighborhood, we then calculate the probability that each household chooses each housing type, and aggregating these choices to the neighborhood level, calculate the predicted sociodemographic composition of each neighborhood. We then replace the initial neighborhood sociodemographic measures with these new measures and start the loop again – i.e., calculate a new set of market clearing prices with these updated neighborhood sociodemographic measures. We continue this process until the neighborhood sociodemographic measures converge. The set of household location decisions corresponding to these new measures along with the vector of market clearing housing prices describe the new equilibrium.[36]

As discussed in Section 3, uniqueness is not a generic property of our sorting model. Without this property, it is sometimes difficult to justify counterfactual simulations corresponding to non-marginal changes in the primitives of the underlying model. As argued by Debreu (1969), however, the property of local uniqueness provides a coherent basis for conducting counterfactual simulations associated with a *marginal* change in the model's primitives. In this case, the results of our equilibrium counterfactual simulations correspond to a series of GE comparative static measures estimated at the current equilibrium. In general, one can verify whether the actual equilibrium is locally unique by checking that the derivative of the implicit function mapping that defines the vector of equilibrium prices has a non-zero

[36] It is also important to point out that because the model itself does not perfectly predict the housing choices that individuals make, the neighborhood sociodemographic measures initially predicted by model, $Z_n^{PREDICT}$, will not match the actual sociodemographic characteristics of each neighborhood, Z_n^{ACTUAL}. Consequently, before calculating the new equilibrium for any simulation, we first solve for the initial prediction error associated with each

determinant at the current equilibrium. In the results that follow, we present the results of a counterfactual simulation that increases the income of households in the top income quartile by 10 percent in order to ensure that the reported results are not sensitive to errors related to rounding and the convergence criterion used in the computation. Such a change is certainly small enough to be considered marginal and the new equilibrium appears to have a very similar structure to the existing one.

An Increase in Income Inequality. We now present the results of this counterfactual simulation, starting with an examination of the impact of this change on neighborhood stratification. Table 4 presents a series of exposure measures that describe the average neighborhood composition (in terms of income) for households in each income quartile before and after the simulation. The measures shown in the first row imply, for instance, that households in the bottom income quartile live in Census block groups that have on average 32.4 percent of households in the bottom income quartile as opposed to 16.5 percent in the top quartile. Households in the top income quartile, on the other hand, live in neighborhoods that have on average 16.6 percent of households in the bottom income quartile and 37 percent in the top quartile. The lower panel of Table 4 reports analogous exposure rates calculated using the location decisions predicted in counterfactual equilibrium. As one would expect, an increase in the income of the top income quartile leads to additional income stratification, leading to a 2 percentage-point (5 percent) increase in the exposure of households in the top income quartile to one another.

---

neighborhood n: $\omega_n = Z_n^{PREDICT} - Z_n^{PREDICT}$. We add this initial prediction error ω_n to the sociodemographic measures calculated in each iteration before substituting these measures back into the utility function.

Table 5 reports a number of consumption measures before and after the simulation, describing the consequences of an increase in the income for the top quartile for the consumption of neighborhood and housing amenities for households at all points of the income distribution. The rows of the table report the average monthly house price, home-ownership rate, average commuting distance, and the average consumption of house size, school quality, crime, neighborhood income and education for each income quartile. Looking first at the total amount households in each quartile spend on housing, notice that households in the top income quartile increase their spending on housing by slightly more than the 10 percent increase in their incomes (possible if housing and neighborhood amenities are luxury goods on the margin). The increased spending power of households in the top income quartile also affects the consumption of households in the other income quartiles, but by much smaller percentages, from 2 percent for those in the bottom quartile to just over 5 percent for those in the third quartile.

The average income of the neighbors of households in the top income quartile also increases by nearly 10 percent in the new equilibrium, combining the effect of the 10 percent increase in income of the top quartile households that these households were already exposed and the increased income stratification shown in Table 4. Interestingly, the corresponding increases in the consumption of housing and neighborhood attributes by households in the top income quartile are much smaller in percentage terms, averaging around only a 1-4 percent improvement for home-ownership, house size, the crime rate, and college-educated neighbors. In this way, in competing for housing and neighborhood attributes in fixed supply, much of the increased spending power of the top income quartile is competed away in bidding for the best houses and neighborhoods available in the market.

This competitive bidding also affects households in the other income quartiles, having an especially negative effect on households in the third quartile. That the consumption of housing and neighborhood attributes drops most markedly for the third income quartile is intuitive as these households are most directly in competition for the types of houses and neighborhoods in which households in the top income quartile reside. As Table 5 makes clear, these households wind up paying approximately 5 percent more for housing, while experiencing *lower* levels of consumption for all of the neighborhood and housing attributes shown except for the average income of their neighbors, which increases by 4 percent. The effect of an increase in the income of households in the top income quartile on the consumption of housing and neighborhood amenities by households in the poorest income quartile is decidedly less marked.

The changes in consumption reported in Table 5 suggest that the implicit prices of various housing and neighborhood amenities may be changing considerably for households near the top of the income distribution as the now richer top quartile households bid up housing prices for the most desirable houses and neighborhoods. Table 6 reports the results of selected coefficients for six weighted hedonic price regressions both before and after the simulation. In each case, housing price is regressed on a series of housing and neighborhood attributes, with weights determined by how close the income of the occupant is to the 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, and 95^{th} percentile of the income distribution, respectively.[37] Thus the first regression provides an indication of the implicit price that households near the 10^{th} percentile of the income distribution for various housing and neighborhood amenities, for example.

Comparing the equilibrium after the simulation to that before, the implicit prices of housing and neighborhood attributes are affected throughout the distribution, although slightly

[37] The particular form of the weight that we use is given by $w_i = {}^{10,000}\!/_{10,000+|inc_i - inc(p)|}$ for the pth percentile.

larger effects occur near the top of the income distribution. Consequently, the increased spending power of high-income households has an affect on prices throughout the housing market, thereby dampening the potential benefit of this income increase. The households most negatively affected by the change, however, are those with income near the top quartile but who do not get the 10 percent increase. These households essentially face the greatest increase in implicit prices without any additional spending power and, as Table 5 reveals, consume lower levels of housing and neighborhood amenities as a result.

7 CONCLUSION

This paper introduces an equilibrium model of residential sorting designed to make full use of newly available Census microdata that provide residential and employment locations down to the level of a city block. This equilibrium model permits a more flexible characterization of preferences than has been possible in the equilibrium sorting models that have been taken to the data thus far, allowing sorting over many housing and neighborhood attributes and bringing the geography of the urban housing market into the model in a natural way. Using a sample of almost quarter of a million households and their corresponding houses and neighborhoods for the San Francisco Bay Area, we estimate a rich set of household preferences for housing and neighborhood attributes, accounting for important endogeneity problems that arise due to the correlation of unobserved aspects of housing and neighborhood quality with equilibrium housing prices and neighborhood sociodemographic compositions, correlations that are induced by residential sorting.

The estimated preference parameters imply that commuting distance, school quality, crime, housing attributes, and particularly neighborhood sociodemographic composition all play

a significant role in the typical household's location decision. Conditional on income, the estimates reveal that college-educated households have clear preferences for living with like households and that households of each race strongly prefer neighborhoods in which a sizeable fraction of their neighbors are of the same race. Demand for desirable housing and neighborhood attributes tends to be an increasing function of a number of household characteristics that proxy to some degree for lifetime wealth, including income from capital and non-capital sources, current employment, and educational attainment.

Using the estimated preference parameters, we conduct a counterfactual simulation that illustrates the capabilities of the model as a tool for analyzing economic or policy changes accounting for effects on residential sorting and housing prices throughout the market. In particular, we characterize a new sorting equilibrium following a 10 percent increase in the income of the richest 25 percent of households in the metro area. The results of this simulation illustrate how the increased income of the top quartile filters through the housing market, raising the prices of the most desirable houses and neighborhoods in the metropolitan area, in turn affecting the consumption of neighborhood and housing attributes by households throughout the income distribution. The price increases induced by increased competition for the most desirable houses and neighborhoods have the effect of eliminating some of the benefit of increased income for top quartile households and has an especially negative welfare effect on households near but not in the top quartile of the income distribution.

REFERENCES

Anas, Alex, (1982), *Residential Location Markets and Urban Transportation: Economic Theory, Econometrics and Public Policy Analysis,* Academic Press, New York.

Bajari, Patrick and Lanier Benkhard, (2002), "Demand Estimation with Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach," mimeo, Stanford University.

Bajari, Patrick, and Matthew Kahn (2001), "Why Do Blacks Live in Cities and Whites Live in Suburbs?" mimeo, Stanford University.

Bartik, Timothy, (1987), "The Estimation of Demand Parameters in Hedonic Price Models," *Journal of Political Economy*, 95:81-88.

Bayer, Patrick, (1999), "Essays Aimed at Understanding Observed Differences in the Consumption of School Quality," Stanford University Dissertation.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan (2003), "A Unified Framework for Measuring Preferences for Schools and Neighborhoods," Economic Growth Center, Yale University Working Paper No. 872.

Bayer, Patrick, Robert McMillan, and Kim Rueben, (2003), "An Equilibrium Model of an Urban Housing Market: A Study of the Causes and Consequences of Residential Segregation," Economic Growth Center, Yale University Working Paper No. 860.

Bayer, Patrick, Robert McMillan, and Kim Rueben, (2004), "Residential Segregation in General Equilibrium," Economic Growth Center, Yale University Working Paper No. 885.

Bayer, Patrick and Christopher Timmins (2003), "On the Equilibrium Properties of Locational Sorting Models," Economic Growth Center, Yale University Working Paper No. 861.

Benabou, Roland, (1993), "The Workings of a City: Location, Education, and Production," *Quarterly Journal of Economics*, 108(3), pp.619-652.

Benabou, Roland, (1996), "Heterogeneity, Stratification, and Growth: Macroeconomic Implications of Community Structure and School Finance," *American Economic Review*, Vol. 86, No. 3., pp. 584-609.

Berry, Steven, (1994), "Estimating Discrete-Choice Models of Product Differentiation," *RAND Journal of Economics*, Vol. 25, pp. 242-262.

Berry, Steven, James Levinsohn, and Ariel Pakes, (1995), "Automobile Prices in Market Equilibrium," *Econometrica*, Vol 63, pp. 841-890.

Berry, Steven, Oliver Linton, and Ariel Pakes, (2002), "Limit Theorems for Estimating the Parameters of Differentiated Product Demand Systems," mimeo, Yale University.

Black, Sandra (1999) "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics*, May 1999.

Brock, William A., and Durlauf, Steven N. (2001), "Discrete Choice with Social Interactions." *Review of Economic Studies*, Vol. 68 April, pp. 235-60.

Brock, William A., and Durlauf, Steven N. (2003), "Multinomial Choice with Social Interactions." *NBER Working Paper T0288*.

Brown, James and Harvey Rosen (1982), "On the Estimation of Structural Hedonic Price Models," *Econometrica*, 50: 765-9.

Coffey, Bentley (2003), "A Reexamination of Air Pollution's Effects on Infant Health: Dos Mobility Matter?" mimeo, Duke University.

Debreu, Gerard (1969), "Economies with a Finite Set of Equilibria," Irving Fisher lecture at the Brussels meeting of the Econometric Society, September, reprinted in *Mathematical Economics: Twenty Papers of Gerard Debreau*, Cambridge University Press, 1983.

Ekeland, Ivar, James Heckman, and Lars Nesheim, (2002), "Identification and Estimation of Hedonic Models," mimeo, University of Chicago.

Epple, Dennis, (1987), "Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products," *Journal of Political Economy*, 107: 645-81.

Epple, D., R. Filimon, and T. Romer, (1984), "Equilibrium Among Local Jurisdictions: Towards an Integrated Approach of Voting and Residential Choice," *Journal of Public Economics*, 24, pp. 281-304.

Epple, D., R. Filimon, and T. Romer, (1993), "Existence of Voting and Housing Equilibrium in a System of Communities with Property Taxes," *Regional Science and Urban Economics*, 23, pp. 585-610.

Epple, Dennis and Holger Sieg, (1999), "Estimating Equilibrium Models of Local Jurisdictions," *Journal of Political Economy*, Vol. 107, No. 4., pp. 645-681.

Epple, Dennis, Thomas Romer and Holger Sieg, (2001), "Interjurisdictional Sorting and Majority Rule: An Empirical Analysis," *Econometrica*, Vol. 69, No. 6., pp. 1437-1455.

Fernandez, Raquel and Richard Rogerson, (1996), "Income Distribution, Communities, and the Quality of Public Education." *Quarterly Journal of Economics*, Vol. 111, No. 1., pp. 135-164.

Fernandez, Raquel and Richard Rogerson, (1998), "Public Education and Income Distribution: A Dynamic Quantitative Evaluation of Education Finance Reform " *American Economic Review*, 88:4., pp. 813-33.

Ferreira, Fernando (2003), "You Can Take It with You: Transferability of Proposition 13 Tax benefits, Residential Mobility, and Willingness to Pay for Housing Amenities," mimeo, U of California, Berkeley.

Ferreyra, Maria, (2003), "Estimating the Effects of Private School Vouchers in Multi-District Economies," mimeo, Carnegie-Mellon University.

Gabriel, S. and S. Rosenthal, (1989), "Household Location and Race: Estimates of a Multinomial Logit Model," *Review of Economics and Statistics*, 71: 240-9.

McFadden, Daniel, (1973), "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, eds., *Frontiers of Econometrics*, Academic Press, New York.

McFadden, Daniel, (1978), "Modeling the Choice of Residential Location," in eds. Karlquist, A., et al., Spatial Interaction Theory and Planning Models, Elsevier North-Holland, New York.

Nechyba, Thomas J., (1999), "School Finance Induced Migration and Stratification Patterns: the Impact of Private School Vouchers," *Journal of Public Economic Theory*, Vol. 1.

Nechyba, Thomas J., (2000), "Mobility, Targeting, and Private School Vouchers," *American Economic Review*, Vol. 90(1): 130-46.

Nechyba, Thomas J., and Robert P. Strauss, (1998), "Community Choice and Local Public Services: A Discrete Choice Approach," *Regional Science and Urban Economics*, Vol. 28, 51-73.

Quigley, John M., (1985), "Consumer Choice of Dwelling, Neighborhood, and Public Services," *Regional Science and Urban Economics*, Vol. 15(1).

Rosen, Sherwin, (1974), "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 82: 34-55.

Schelling, Thomas C., (1969), "Models of Segregation." *American Economic Review*, 59(2): 488-93.

Schelling, Thomas C., (1971), "Dynamic Models of Segregation," *Journal of Mathematical Sociology*, 1: 143-186 .

Sieg, Holger, V. Kerry Smith, H. Spencer Banzaf and Randall Walsh, (forthcoming) "Estimating the General Equilibrium Benefits of Large Changes in Spatially Delineated Public Goods," *International Economic Review*.

Tiebout, Charles M., (1956), “A Pure Theory of Local Expenditures,” *Journal of Political Economy*, 64: 416-424.

Timmins, Christopher (2003), "If You Can't Take the Heat, Get Out of the Cerrado... Recovering the Equilibrium Amenity Cost of Non-Marginal Climate Change in Brazil," mimeo, Yale University.

Walsh, Randall (2004), “Endogenous Open Space Amenities in a Locational Equilibrium,” CEA Working Paper No. 04-03, February.

Table 1. Overall Sample and Sub-Sample Near School District Boundaries

Sample	**full sample**		**within 0.25 miles of boundaries**				
Boundary/Weights			**actual sample**	**high test score side***	**low test score side***	**t-test for**	**weighted sample**
Observations	**242,100**		**27,958**	**13,348**	**14,610**	**difference in**	**27,958**
	(1)	(2)	(3)	(4)	(5)	**means**	(6)
	Mean	S.D.	Mean	Mean	Mean	((4) versus (5))	Mean
Housing/Neighborhood Characteristics							
monthly house price	1,087	755	1,130	1,158	1,105	5.71	1,098
average test score	527	74	536	558	515	50.96	529
1 if unit owned	0.597	0.491	0.629	0.632	0.626	1.04	0.616
number of rooms	5.114	1.992	5.170	5.207	5.134	3.13	5.180
1 if built in 1980s	0.143	0.350	0.108	0.118	0.099	5.09	0.148
1 if built in 1960s or 1970s	0.391	0.488	0.424	0.412	0.437	4.22	0.406
elevation	210	179	193	194	192	1.14	212
population density	0.434	0.497	0.352	0.349	0.355	2.08	0.374
crime index	8.184	10.777	6.100	6.000	6.192	2.36	7.000
% Census block group white	0.681	0.232	0.704	0.712	0.686	9.62	0.676
% Census block group black	0.081	0.159	0.071	0.065	0.076	6.21	0.080
% Census block group Hispanic	0.110	0.114	0.113	0.107	0.119	8.62	0.117
% Census block group Asian	0.122	0.120	0.112	0.110	0.113	2.50	0.121
% block group college degree or more	0.438	0.196	0.457	0.463	0.451	5.14	0.433
average block group income	54,744	26,075	57,039	58,771	55,457	10.23	55,262
Household Characteristics							
household income	54,103	50,719	56,663	58,041	55,405	4.20	55,498
1 if children under 18 in household	0.333	0.471	0.324	0.322	0.325	0.54	0.336
1 if black	0.076	0.264	0.066	0.062	0.070	2.69	0.076
1 if Hispanic	0.109	0.312	0.111	0.102	0.119	4.54	0.115
1 if Asian	0.124	0.329	0.112	0.114	0.110	1.06	0.121
1 if white	0.686	0.464	0.706	0.717	0.696	3.86	0.682
1 if college degree or more	0.438	0.497	0.460	0.467	0.454	2.64	0.441
age (years)	47.607	16.619	47.890	48.104	47.699	1.99	47.660
1 if working	0.698	0.459	0.705	0.702	0.709	1.28	0.701
distance to work (miles)	8.843	8.597	8.450	8.412	8.492	0.82	8.490

Notes: Columns 1 and 2 report the mean and standard deviation for key variables for the full sample. Column 3 reports means for the sample of houses within 0.25 miles of a school district boundary. Columns 4 and 5 report means on the high versus low test score side of boundaries. Column 6 provides a t-statistic for a test of whether the means reported in columns 4 and 5 are equal. Column 7 reports weighted means for the sample of houses within 0.25 miles of a school district boundary. Weights are constructed so as to make the boundary sample more representative of the full sample and are described in the main text. In constructing columns 4 and 5, we assign each house in the full sample to the nearest school district boundary, noting whether its local school has a higher test score than the school associated with the closest Census block on the other side of the boundary.

Table 2: Implied Mean MWTP Measures

	Residential Sorting Model		Hedonic Price Regressions	
Sample	within .25 mile of boundaries		within .25 mile of boundaries	
Boundary Fixed Effects	No	Yes	No	Yes
Observations	27,958	27,958	27,958	27,958
	(2)	(3)	(2)	(3)
% Black*	-285.46	-233.94	-94.96	-40.46
	(32.06)	(38.87)	(35.28)	(42.74)
% Hispanic*	-37.19	104.11	106.60	254.31
	(46.83)	(59.01)	(51.54)	(64.88)
% Asian*	-69.84	149.77	-1.69	241.13
	(45.68)	(55.21)	(50.27)	(60.71)
% College Degree or More	185.74	164.78	235.04	177.11
	(25.96)	(39.42)	(28.57)	(43.34)
Average Income*	89.48	85.44	113.26	109.22
	(2.18)	(2.64)	(2.40)	(2.90)
Average Test Score (in s.d.'s)	16.69	21.46	19.01	23.67
	(4.23)	(5.29)	(4.66)	(5.81)
Owner-Occupied	141.08	148.15	117.59	125.63
	(7.40)	(7.38)	(8.14)	(8.12)
Number of Rooms	111.67	109.28	123.91	121.72
	(1.95)	(1.96)	(2.15)	(2.16)
Built in 1980s	71.36	87.40	80.58	108.57
	(9.29)	(10.00)	(10.23)	(10.99)
Built in 1960s or 1970s	1.32	2.48	-4.40	4.87
	(6.86)	(7.47)	(7.55)	(8.21)
Elevation (in s.d.'s)	-26.08	11.02	-25.20	11.65
	(3.90)	(7.14)	(4.29)	(7.85)
Population Density(in s.d.'s)	20.33	11.50	24.44	7.29
	(6.90)	(8.61)	(7.59)	(9.47)
Crime Index (in s.d.'s)	-5.31	-10.96	10.08	3.88
	(7.00)	(17.25)	(7.70)	(18.97)
F-statistic for boundary fixed effects		4.162		8.754

Notes: All neighborhood attributes are measured using the corresponding Census block group. Specifications shown in the table also include controls for interactions between neighborhood racial composition variables and average income as well as land use (% industrial, % residential, % commercial, % open space, % other) in 1, 2, 3, 4, and 5 mile rings around location and six variables that characterize the housing stock in each of these rings.
*Coefficients for % Asian,% Black, % Hispanic, Average Income reported at mean.

Table 3: Heterogeneity in Marginal Willingness to Pay

	Mean	Children Under 18 (vs. none)	Non-Capital Income (+ $10,000)	Capital Income (+10,000)	Black (vs. White)	Hispanic (vs. White)	Asian (vs. White)	Some College vs. HS or les	College Degree vs. HS or les	Working (vs. not)	Age (+10 yrs)	Black* Hhld Income (+10,000)	Hispanic* Hhld Income (+10,000)	Asian* Hhld Income (+10,000)
Number of Rooms	109.3	31.2	4.2	2.5	0.3	-13.7	-32.4	0.9	1.0	0.3	2.0			
(+1 room)	(2.0)	(2.6)	(0.2)	(0.5)	(4.6)	(3.4)	(2.2)	(1.3)	(2.2)	(2.6)	(0.5)			
Owner-Occupied	148.0	-12.1	15.9	26.1	-50.0	-4.9	91.5	1.2	21.3	33.5	50.8			
(vs. Renter Occupied)	(7.4)	(5.6)	(0.8)	(6.1)	(19.4)	(13.6)	(12.0)	(2.6)	(6.1)	(6.2)	(1.8)			
Built in 1980s	87.4	-21.8	7.0	9.9	10.7	-6.0	27.8	12.2	30.5	52.4	-16.9			
(vs. Pre-1960)	(10.0)	(9.8)	(0.9)	(6.9)	(33.0)	(21.5)	(17.3)	(7.8)	(8.2)	(6.7)	(3.9)			
Built in 1960-79	2.5	5.5	1.5	1.7	37.2	-15.0	23.9	7.4	4.9	23.3	-5.1			
(vs. Pre-1960)	(7.5)	(5.1)	(0.8)	(10.1)	(10.0)	(7.9)	(7.7)	(5.5)	(9.2)	(10.8)	(2.7)			
Average Test Score	21.5	6.6	0.1	3.1	-13.1	-4.0	4.5	4.1	11.2	7.7	6.1			
(+1 s.d.)	(5.3)	(4.8)	(0.2)	(0.9)	(7.3)	(6.2)	(3.4)	(1.2)	(2.4)	(2.8)	(1.3)			
Elevation	11.0	4.5	1.1	-1.0	-5.6	-7.1	0.3	3.2	4.9	-1.0	3.3			
(+1 s.d.)	(7.1)	(5.0)	(0.5)	(0.3)	(7.1)	(4.7)	(8.9)	(2.5)	(5.0)	(3.4)	(1.0)			
Population Density	11.5	-25.7	0.9	3.2	-32.2	-1.6	0.2	-0.1	8.7	-13.8	-3.1			
(+1 s.d.)	(8.6)	(5.8)	(0.3)	(1.0)	(10.9)	(8.4)	(8.7)	(1.1)	(2.4)	(6.0)	(1.4)			
Crime Index	-11.0	1.1	-0.9	2.6	28.2	2.3	0.8	-1.0	12.9	-8.9	7.6			
(+1 s.d.)	(17.3)	(2.6)	(0.2)	(0.4)	(4.9)	(4.0)	(5.9)	(2.2)	(4.3)	(4.5)	(1.8)			
% Black*	-23.4	8.5	-0.7	-4.4	85.8	23.0	22.6	-1.2	5.0	-3.9	-0.9	-2.1		
(+10%)	(3.9)	(1.5)	(0.2)	(0.4)	(2.2)	(1.5)	(1.6)	(0.5)	(1.0)	(3.1)	(0.6)	(0.7)		
% Hispanic*	10.4	13.0	0.2	-3.4	35.3	56.6	19.6	-3.8	-4.5	-0.7	-5.0		2.3	
(+10%)	(5.9)	(1.7)	(0.3)	(0.8)	(3.6)	(3.2)	(3.1)	(0.8)	(1.7)	(1.8)	(0.5)		(0.6)	
% Asian*	14.9	8.8	0.3	-3.2	38.3	13.6	87.5	-1.2	-9.8	-2.5	0.6			-1.7
(+10%)	(5.5)	(3.2)	(0.2)	(0.2)	(5.7)	(2.2)	(2.6)	(1.0)	(2.0)	(1.6)	(0.5)			(0.3)
% College Degree or More	16.4	-12.2	0.6	3.1	16.8	2.1	-1.4	4.3	46.8	-9.4	-1.8			
(+10%)	(3.9)	(2.0)	(0.7)	(0.3)	(6.1)	(5.1)	(4.3)	(0.8)	(2.2)	(2.0)	(0.7)			
Average Income*	85.4	2.2	0.9	1.3	-17.8	4.9	-1.1	-0.2	-6.6	2.1	0.7			
(+10,000)	(2.6)	(1.3)	(0.1)	(0.9)	(2.4)	(5.0)	(2.9)	(0.7)	(2.7)	(1.4)	(0.3)			
% Black*Average Income	2.9		0.7		9.7							-0.4		
(+10,000*8.1%)	(2.9)		(0.1)		(1.5)							(0.3)		
% Hispanic*Average Incom	-20.3		0.7			3.6							-1.0	
(+10,000*11.0%)	(2.4)		(0.1)			(0.9)							(0.2)	
% Asian*Average Income	1.4		0.1				1.9							0.0
(+10,000*12.2%)	(1.5)		(0.1)				(0.7)							(0.1)
Distance to Work		1.2	-0.1	-1.7	-1.0	0.7	0.8	-0.1	0.6	-50.0	-0.4			
(+1 mile)		(0.3)	(0.0)	(0.0)	(0.4)	(0.4)	(0.3)	(0.1)	(0.2)	(0.3)	(0.1)			

Notes: The first column reports the estimated mean MWTP for the changes in housing or neighborhood attributes described in the row headings. The remaining columns report the difference in MWTP associated with the comparison of household characteristics shown in the column heading. So, for example, the first entry of the table implies that, on average, households are willing-to-pay $109 more per month on the margin for an additional room, while the second entry in the first row implies that households with children are willing to pay an average of $31 per month more for a room than households without children. *Coefficients for % Asian,% Black, % Hispanic, Average Income reported at mean.

Table 4: Income Stratification: Effects of Increasing Income of Top Quartile by 10%

Panel A: Pre-Simulation

	Average Exposure to Households in Income Quartile			
	1	**2**	**3**	**4**
Income Quartile 1	0.324	0.275	0.235	0.165
Income Quartile 2	0.278	0.265	0.250	0.207
Income Quartile 3	0.236	0.248	0.258	0.257
Income Quartile 4	0.166	0.207	0.258	0.370

Panel B: Post-Simulation

	Average Exposure to Households in Income Quartile			
	1	**2**	**3**	**4**
Income Quartile 1	0.327	0.278	0.237	0.157
Income Quartile 2	0.281	0.267	0.251	0.200
Income Quartile 3	0.239	0.250	0.259	0.252
Income Quartile 4	0.158	0.200	0.253	0.390

Note: Each entry in the table shows the average exposure of households in the income quartile shown in the row heading to households in the income quartile shown in the column heading. Numbers are reported for the sample and for a counterfactual simulation that increases the income of households in the top income quartile by 10 percent.

Table 5: Housing/Neighborhood Consumption: Effects of Increasing Income of Top Quartile by 10%

	Average Consumption By Households in Income Quartile			
	1	**2**	**3**	**4**
	Average Monthly Rental Value			
Pre-Simulation	726	903	1112	1608
Post-Simulation	734	926	1161	1790
	Ownership Rates			
Pre-Simulation	0.37	0.50	0.67	0.85
Post-Simulation	0.38	0.50	0.66	0.86
	House Size			
Pre-Simulation	3.95	4.61	5.41	6.50
Post-Simulation	3.95	4.61	5.39	6.52
	Average Test Score			
Pre-Simulation	502	517	529	559
Post-Simulation	505	517	528	557
	Average Crime Rate			
Pre-Simulation	12.36	8.65	6.67	5.05
Post-Simulation	12.29	8.72	6.76	4.94
	Average Commute			
Pre-Simulation	6.32	8.12	9.56	9.98
Post-Simulation	7.09	8.87	10.30	10.81
	Average Neighborhood Income (,000s)			
Pre-Simulation	43.3	48.5	54.6	69.6
Post-Simulation	43.9	50.0	56.5	76.0
	Percent College Educated			
Pre-Simulation	0.36	0.41	0.45	0.54
Post-Simulation	0.34	0.40	0.45	0.56

Note: This table reports the consumption of housing and neighborhood amenities by households of each race in each quartile of the overall income distribution. Numbers are reported for the sample and for a counterfactual simulation that increases the income of households in the top income quartile by 10 percent.

Table 6: Changes in Implicit Prices: Effects of Increasing Income of Top Quartile by 10% for Select Characteristics

PRE-SIMULATION	**Coefficients from Hedonic Price Regressions Weighted to Center on Income Percentile**					
	10	**25**	**50**	**75**	**90**	**95**
Owner-Occupied	90.7	95.3	112.3	149.8	165.5	160.1
Number of Rooms	103.1	103.9	107.3	114.2	125.8	132.5
Built in 1980s	97.7	94.4	89.7	96.0	135.7	161.8
Average Test Score (in s.d.s)	27.2	27.8	29.9	30.4	30.4	34.7
Average Income (/10,000)	87.6	84.7	82.3	83.2	90.7	97.5

POST-SIMULATION	**Coefficients from Hedonic Price Regressions Weighted to Center on Income Percentile**					
	10	**25**	**50**	**75**	**90**	**95**
Owner-Occupied	105.2	109.9	130.0	170.3	183.8	176.3
Number of Rooms	111.9	113.0	117.1	125.5	138.4	145.4
Built in 1980s	81.2	77.8	72.6	84.6	132.3	160.8
Average Test Score (in s.d.s)	30.1	30.1	31.9	32.5	33.5	38.5
Average Income (/10,000)	95.3	92.9	91.1	92.3	100.7	108.1

DIFFERENCE						
	10	**25**	**50**	**75**	**90**	**95**
Owner-Occupied	14.5	14.6	17.7	20.6	18.4	16.3
Number of Rooms	8.8	9.2	9.8	11.3	12.6	12.9
Built in 1980s	-16.5	-16.6	-17.0	-11.4	-3.4	-0.9
Average Test Score (in s.d.s)	3.0	2.3	2.0	2.1	3.1	3.8
Average Income (/10,000)	7.8	8.2	8.7	9.1	10.0	10.6

Note: This table reports the selected coefficients for six weighted price regressions estimated pre- and post-simulation, respectively. The weight in each case depends on income of the occupant of the house and is given by 10,000/(10,000+|income - income-pth percentile|) for the regression associated with p-th percentile, which puts more weight on households near the p-th percentile. Results are reported for the sample and for a counterfactual simulation that increases the income of households in the top income quartile by 10 percent.

Appendix Table 1: Interaction Parameter Estimates

	Household Characteristic												
	Hhld Total Income	Children Under 18	Black	Hispanic	Asian	Some College	College Degree or More	Working	Age	Hhld Capital Income	Black* Hhld Income	Hispanic* Hhld Income	Asian* Hhld Income
Housing/Neighborhood Attribute													
Monthly House Price	0.371	0.074	0.025	-0.076	0.067	0.118	0.198	0.091	0.119	0.035	0.041	0.151	0.076
	(0.016)	(0.028)	(0.058)	(0.057)	(0.049)	(0.028)	(0.030)	(0.030)	(0.031)	(0.025)	(0.062)	(0.050)	(0.058)
Owner-Occupied	0.739	-0.052	-0.121	-0.014	0.274	0.011	0.095	0.140	0.776	0.261			
	(0.036)	(0.024)	(0.047)	(0.040)	(0.036)	(0.024)	(0.027)	(0.026)	(0.027)	(0.061)			
Number of Rooms	0.785	0.544	0.003	-0.163	-0.394	0.035	0.018	0.005	0.122	-0.167			
	(0.045)	(0.046)	(0.045)	(0.040)	(0.027)	(0.047)	(0.039)	(0.044)	(0.028)	(0.030)			
Built in 1980s	0.233	-0.067	0.018	-0.013	0.059	0.080	0.097	0.156	-0.184	0.053			
	(0.029)	(0.030)	(0.057)	(0.045)	(0.037)	(0.051)	(0.026)	(0.020)	(0.042)	(0.037)			
Built in 1960-79	0.070	0.024	0.090	-0.044	0.071	0.068	0.022	0.097	-0.077	0.005			
	(0.035)	(0.022)	(0.024)	(0.023)	(0.023)	(0.050)	(0.041)	(0.045)	(0.041)	(0.029)			
Average Test Score	0.007	0.058	-0.065	-0.024	0.028	0.077	0.102	0.065	0.191	0.155			
	(0.021)	(0.042)	(0.036)	(0.037)	(0.021)	(0.023)	(0.022)	(0.024)	(0.040)	(0.045)			
Elevation	0.100	0.039	-0.027	-0.042	0.002	0.059	0.044	-0.008	0.104	-0.108			
	(0.045)	(0.044)	(0.035)	(0.028)	(0.054)	(0.047)	(0.045)	(0.029)	(0.030)	(0.029)			
Population Density	0.087	-0.225	-0.159	-0.009	0.001	-0.003	0.079	-0.118	-0.096	0.116			
	(0.030)	(0.051)	(0.054)	(0.050)	(0.053)	(0.020)	(0.022)	(0.051)	(0.043)	(0.037)			
Crime Index	-0.083	0.010	0.139	0.014	0.005	-0.018	0.117	-0.076	0.235	0.181			
	(0.023)	(0.023)	(0.024)	(0.024)	(0.036)	(0.041)	(0.039)	(0.038)	(0.055)	(0.030)			
% Black	-0.382	0.119	0.482	0.218	0.219	-0.037	0.072	-0.053	-0.043	-0.304	-0.043		
	(0.060)	(0.041)	(0.035)	(0.028)	(0.031)	(0.029)	(0.029)	(0.085)	(0.060)	(0.052)	(0.051)	.	
% Hispanic	-0.329	0.130	0.198	0.279	0.137	-0.080	-0.047	-0.007	-0.177	-0.211		0.232	
	(0.061)	(0.033)	(0.041)	(0.043)	(0.043)	(0.032)	(0.035)	(0.034)	(0.033)	(0.099)		(0.047)	
% Asian	-0.014	0.092	0.227	0.098	0.550	-0.027	-0.107	-0.026	0.023	-0.219			-0.080
	(0.050)	(0.067)	(0.068)	(0.032)	(0.038)	(0.045)	(0.044)	(0.033)	(0.038)	(0.034)			(0.036)
% College Degree	0.115	-0.209	0.163	0.025	-0.017	0.156	0.834	-0.157	-0.108	0.258			
	(0.137)	(0.035)	(0.059)	(0.060)	(0.052)	(0.028)	(0.040)	(0.033)	(0.045)	(0.028)			
Average Income	0.232	0.050	-0.229	0.077	-0.017	-0.012	-0.155	0.046	0.056	0.056			
	(0.031)	(0.030)	(0.031)	(0.078)	(0.046)	(0.036)	(0.064)	(0.032)	(0.023)	(0.039)			
% Black *Average Income	0.280		0.191								-0.034		
	(0.038)		(0.029)								(0.024)		
% Hispanic *Average Income	0.347			0.106								-0.150	
	(0.025)			(0.025)								(0.025)	
% Asian *Average Income	0.053				0.091								-0.008
	(0.063)				(0.034)								(0.042)
Distance to Work	-0.114	0.162	-0.077	0.059	0.071	-0.039	0.080	-6.380	-0.167	-1.294			
	(0.035)	(0.037)	(0.030)	(0.036)	(0.030)	(0.029)	(0.031)	(0.032)	(0.031)	(0.031)			

Note: Parameter estimates reported with all variables normalized to have mean zero, standard deviation one. Standard errors are in parentheses.

APPENDICES FOR "AN EQUILIBRIUM MODEL OF SORTING IN AN URBAN HOUSING MARKET" by Patrick Bayer, Robert McMillan, and Kim Rueben

This document contains three appendices for the paper "An Equilibrium Model of Sorting in an Urban Housing Market." A *Data Appendix* documents the sources for the data and the construction of variables used in the analysis. A *Theory Appendix* provides proofs for the propositions and lemmas in Section 3 of the paper. A *Results Appendix* relates the main parameter estimates presented in the paper to analogous estimates from a series of hedonic price regressions.

DATA APPENDIX

1. Census Variables

House Prices. Because house values are self-reported, it is difficult to ascertain whether these prices represent the current market value of the property, especially if the owner purchased the house many years earlier. Fortunately, the Census contains other information that helps us to examine this issue and correct house values accordingly. In particular, the Census asks owners to report a continuous measure of their annual property tax payment. The rules associated with Proposition 13 imply that the vast majority of property tax payments in California should represent exactly 1 percent of the transaction price of the house at the time the current owner bought the property or the value of the house in 1978. Thus, by combining information about property tax payments and the year that the owner bought the house (also provided in the Census in relatively small ranges), we are able to construct a measure of the rate of appreciation implied by each household's self-reported house value. We use this information to modify house values for those individuals who report values much closer to the original transaction price rather than current market value. In our study most households list the purchase price of their house rather than an estimated market value for their house. Thus if two identical houses were found in the census data but one was last sold in 1989 and one was last sold in 1969 we find on average the listed market price of the more recently sold house is on average 15 percent higher than the other house.

A second deficiency of the house values reported in the Census is that they are top-coded at $500,000, a top-code that is often binding in California. Again, because the property tax payment variable is continuous and not top-coded, it provides information useful in distinguishing the values of the upper tail of the value distribution. We find that top-coding was fairly predominant in the Bay Area and that higher top-codes may be useful to gain a better understanding of house prices in expensive markets like California or New York.

The exact procedure that we use to adjust self-reported house values is as follows. We first regress the log of self-reported house value on the log of the estimated transaction price (100 times the property tax payment), and a series of dummy variables that characterize the tenure of the current owner:

$$\text{(A1)} \qquad \log(V_j) = \alpha_1 \log(T_j) + \alpha_2 y_j + \omega_j$$

where V_j represents the self-reported house value, T_j represents the estimated transaction price, and y_j represents a series of dummy variables for the year that the owner bought the house. If owner-estimated house values were indeed current market values and houses were identical except for owner tenure, this regression would return an estimate of 1 for α_1 and the estimated α_2 coefficients would indicate the appreciation of house values in the Bay Area over the full period of analysis. If owners tend to underreport house values, especially when they have lived in the house for a long time, the estimated α_2 parameters will likewise underreport appreciation in the market. In this way, the estimated α_2 parameters represent a conservative estimate of appreciation. Given the estimates of equation (2), we construct a predicted house value for each house in the sample and replace the owner-reported value with this measure when this predicted measure exceeds the owner-reported value. In practice, in order to allow for different rates of appreciation in different regions of the housing market, we conduct these regressions separately for each of the 45 Census PUMA (areas with at least 100,000 people) in our sample and allow appreciation to vary with a small set of house characteristics within each PUMA. In this way, the first adjustment that we make to house prices is to adjust owner-reported values for likely under-reporting.

The adjustment to top-coded house prices uses the same approach, using the information on property taxes that are continuous and not top-coded. Using estimates of equation (2) based on a sample of

houses that does not include the top-coded house values, we construct predicted house values for all top-coded houses. This allows us to assign continuous house values for top-coded measures.

Reported Rental Value. We next examined questions of reported monthly rents. While rents are presumably not subject to the same degree of misreporting as house values, it is still the case that renters who have occupied a unit for a long period of time generally receive some form of tenure discount. In some cases, this tenure discount may arise from explicit rent control, but implicit tenure discounts generally occur in rental markets even when the property is not subject to formal rent control. Thus while, this will not lead to errors in the answering of the listed census question it may lead to an inaccurate comparison of rents faced by households if they needed to move. In order to get a more accurate measure of the market rent for each rental unit, we utilize a series of locally based hedonic price regressions in order to estimate the discount associated with different durations of tenure in each of over 40 sub-regions within the Bay Area.

In order to get a better estimate of market rents for each renter-occupied unit in our sample, we regress the log of reported rent R_j on a series of dummy variables that characterize the tenure of the current renter, y_j, as well as a series of variables that characterize other features of the house and neighborhood X_j:

$$\text{(A2)} \qquad \log(R_j) = \beta_1 y_j + \beta_2 X_j + \mu_j$$

again running these regressions separately for each of the 45 PUMAs in our sample. To the extent that the additional house and neighborhood variables included in equation (3) control for differences between the stock of rental units with long-term vs. short-term tenants, the β_1 parameters provide an estimate of the tenure discount in each PUMA.[1] In order to construct estimates of market rents for each rental unit in our sample, then, we inflate rents based on the length of time that the household has occupied the unit using the estimates of β_1 from equation (2). In this way, these three price adjustments bring the measures for rents and house values reported in the Census reasonably close to market rates.

Calculating Cost Per Unit of Housing Across Tenure Status. Finally, in order to make owner- and renter-occupied housing prices comparable in our analysis we need to calculate a current rental value for housing. Because house prices reflect the expectations about the future rents for the property they incorporate beliefs about future housing appreciation. To appropriately deflate housing values – and especially to control for differences in expectations about appreciation in different segments of the Bay Area housing market – we regress the log of house price (whether monthly rent or house value) Π_j on an indicator for whether the housing unit is owner-occupied o_j and a series of additional controls for features of the house including the number of rooms, number of bedrooms, types of structure (single-family detached, unit in various sized buildings, etc.), and age of the housing structure as well as a series of neighborhood controls X_j:

$$\text{(A3)} \qquad \log(\Pi_j) = \gamma_1 o_j + \gamma_2 X_j + \eta_j$$

We estimate these hedonic price regressions for each of 40 sub-regions (Census Public Use Microdata Areas - PUMAs) of the Bay Area housing market. These regressions return an estimate of the ratio of house values to rents for each of these sub-regions and we use these ratios to convert house values to a measure of current monthly rent.

2. External Data

We next discuss the additional variables we have added to the Census data to provide a more nuanced understanding of the neighborhood characteristics that affect house prices and residential location decisions. These data sets are linked to census blocks and can be used to determine the appropriateness of the questions and sampling techniques used. This additional data includes:

[1] Interestingly, while we estimate tenure discounts in all PUMAs, the estimated tenure discounts are substantially greater for rental units in San Francisco and Berkeley, the two largest jurisdictions in the Bay Area that had formal rent control in 1990.

School and School District Data. The Teale data center in California provided a crosswalk that matches all Census blocks in California to the corresponding public school district. We have further matched Census blocks to particular schools using a variety of procedures that takes account of the location (at the block level) of each Census block within a school district and the precise location of schools within the district using information on location from the Department of Education. Other school information in these data include:

- 1992-93 CLAS dataset provides detailed information about school performance and peer group measures. The CLAS was a test administered in the early 1990s that will give us information on student performance in math, literature and writing for grades 4, 8 and 10. This dataset presents information on student characteristics and grades for students at each school overall and across different classifications of students, including by race and education of parents.
- 1991-2 CBEDS (California Board of Education data sets) datasets including information from the SIF (school information form) which includes information on the ethnic/racial and gender make-up of students, PAIF – which is a teacher based form that provides detailed information about teacher experience, education and certification backgrounds and information on the classes each teacher teaches, and (LEP census) a language census that provides information on the languages spoken by limited-English speaking students.

Procedures for Assigning School Data. While we have an exact assignment of Census blocks to school districts, we have only been able to attain precise maps that describe the way that city blocks are assigned to schools in 1990 for Alameda County. In the absence of information about within-district school attendance areas, we employ the alternative approaches for linking each house to a school. The crudest procedure assigns average school district characteristics to every house falling in the school district. A refinement on this makes use of distance-weighted averages. For a house in a given Census block, we calculate the distance between that Census block and each school in the school district. We have detailed information characterizing each school and construct weighted averages of each school characteristic, weighting by the reciprocal of the distance-squared as well as enrollment.

As a third approach we simply assign each house to the closest school within the appropriate school district. Our preferred approach (which we use for the results reported in the paper) refines this closest-school assignment by using information about individual children living in each Census block - their age and whether they are enrolled in public school. In particular, we modify the closest-school assignment technique by attempting to match the observed fourth grade enrollment for every school in every school district in the Bay Area. Adjusting for the sampling implicit in the long form of the Census, the 'true' assignment of houses to schools must give rise to the overall fourth grade enrollments observed in the data.

These aggregate numbers provide the basis for the following intuitive procedure: we begin by calculating the five closest schools to each Census block. As an initial assignment, each Census block and all the fourth graders in it are assigned to the closest school. We then calculate the total predicted enrollment in each school, and compare this with the actual enrollment. If a school has excess demand, we reassign Census blocks out of its catchment area, while if a school has excess supply, we expand the school's catchment area to include more districts.

To carry out this adjustment, we rank schools on the basis of the (absolute value of) their prediction error, dealing with the schools that have the greatest excess demand/supply first. If the school has excess demand, we reassign the Census block that has the closest second school (recalling that we record the five closest schools to each Census block, in order), as long as that second school has excess supply. If a school has excess supply, we reassign to it the closest school district currently assigned to a school with excess demand. We make gradual adjustments, reassigning one Census block from each school in disequilibrium each iteration. This gradual adjustment of assignments of Census blocks to schools continues until we have 'market clearing' (within a certain tolerance) for each school. Our actual algorithm converges quickly and produces plausible adjustments to the initial, closest-school assignment.

Land use. Information on land use/land cover digital data is collected by USGS and converted to ARC/INFO by the EPA available at: http://www.epa.gov/ost/basins/ for 1988. We have calculated for each Census block, the percentage of land in a 1/4, 1/2,1, 2, 3, 4 and 5-mile radii that is used for commercial,

residential, industrial, forest (including parks), water (lakes, beaches, reservoirs), urban (mixed urban or built up), transportation (roads, railroad tracks, utilities) and other uses.

Crime data. Information on crime was drawn from the rankings of zipcodes on a scale of 1-10 on the risk of violent crime (homicide, rape or robbery). A score of 5 is the average risk of violent crime and a score of 1 indicates a risk 1/5 the national average and a 10 is 10 or more times the national average. These ratings are provided by CAP index and were downloaded from APBNews.com.

Geography and Topography. The Teale data center in California provided information on the elevation, latitude and longitude of each Census block.

THEORY APPENDIX

Proof of Proposition 1:[2] Following the assumptions of Proposition 1, consider a utility specification that is a linear, decreasing function of p_h:

$$\text{(A4)} \qquad V_h^i = W_h^i(Z^i, X_h, \boldsymbol{x}_h) - \boldsymbol{a}_p^i p_h + \boldsymbol{e}_h^i = W_h^i - \boldsymbol{a}_p^i p_h + \boldsymbol{e}_h^i$$

If **e** is drawn from a continuous distribution, the probability P^i_h that household *i* chooses housing type *h* as:

$$\text{(A5)} \qquad P_h^i = f_h(z^i, \mathbf{Z}, \mathbf{X}, \mathbf{p}, \mathbf{x})$$

is continuous and differentiable in *p* with derivatives that obey the following strict inequalities: $\partial P_h^i / \partial p_h < 0$ and $\partial P_h^i / \partial p_k > 0,\ k \neq h$, if $-\boldsymbol{a}_p^i$ is negative for each household *i*. Aggregating these probabilities over all observed households yields the predicted demand for each housing type *h*, D_h:

$$\text{(A6)} \qquad D_h = \sum_i P_h^i \ .$$

Given the properties of P_h^i just described, D_h is also continuous and differentiable in **p** with derivatives that obey the following strict inequalities: $\partial D_h / \partial p_h < 0$ and $\partial D_h / \partial p_k > 0, k \neq h$. In order for the housing market to clear, the demand for houses of type *h* must equal the supply of such houses and so:

$$\text{(A7)} \qquad D_h = S_h, \quad \forall h \quad \Rightarrow \quad \sum_i P_h^i = S_h \quad \forall h .$$

Also note that for any finite values of $\{p_k, k \neq h\}$, $\hat{D}_h$ approaches arbitrarily close to zero as p_h goes to +∞, while $\hat{D}_h$ approach arbitrarily close to $\sum_h S_h$ as p_h approaches –∞.

Holding the price of one house fixed (without loss of generality set $p_0 = 0$), we will show that a unique vector of prices clears the market, i.e., that a unique vector $\mathbf{p} = D^{-1}(\mathbf{S})$ exists. We begin by defining the element-by-element inverse $r_h(\mathbf{p}, D_h)$. This function is defined as the price of house *h* such that the predicted value D_h exactly equals S_h. That is, **r** is implicitly defined as:

$$\text{(A8)} \qquad D_h(p_1, p_2, \ldots, r_h(\mathbf{p}, D_h), \ldots, p_H) = S_h \quad \forall h$$

Given the properties of the function D_h defined in (A6), this element-by-element inverse exists and is continuous and differentiable in **p**. Note that r_h is strictly increasing in p_k and does not depend on p_h. Also define the vector values $\mathbf{r} = (r_1, \ldots, r_N)$.

---

[2] This proof follows directly the structure of the proof that appears in the technical appendix of Berry (1994). We simply modify it here for our problem.

The element-by-element inverse allows us to transform the problem of solving for the vector inverse into a fixed-point problem, for a vector $\mathbf{p}$ satisfies equation (A7) if and only if $\mathbf{p} = \mathbf{r(p, D)}$. The method of proof is to use a slight variant of Brouwer's fixed-point theorem to prove existence of a fixed point of the element-by-element inverse. It is then necessary to show that there cannot be two such fixed points.

To establish existence, first hold $p_0 = 0$ and note that $r_h(\mathbf{p}, D_h)$ has an upper bound. This upper bound is $r_h(\mathbf{p}', D_h)$ with $\mathbf{p}'$ set equal to any vector in $\mathbf{R}^{\mathbf{N+1}}$ such that $p_k = +\infty$ for $k \neq (h,0)$. Define $\overline{p}$ as the largest values across houses of these upper bounds. There is no lower bound for p_h, but the following lemma allows one to establish existence in the absence of a lower bound.

Lemma. There is a value $\underline{p}$, with the property that if one element of $\mathbf{p}$, say p_h, is lower than $\underline{p}$, then there is a house k such that $r_k(\mathbf{p}, \hat{\mathbf{D}}) > p_k$.

Proof of Lemma. To construct $\underline{p}$, again set $p_k = +\infty$, $\forall\, k \neq (h,0)$. Then define $\underline{p}_h$ as the value of p_h that sets $\hat{D}_0 = S_0$. Define $\underline{p}$ as any value lower than the minimum of the $\underline{p}_h$. Now, if for the vector $\mathbf{p}$ there is an element h such that $p_h < \underline{p}$, then $\hat{D}_0(\mathbf{p}) > S_0$, which implies $\sum_{h=1}^{N} \hat{D}_h(\mathbf{p}) < (N-1) \bullet S_h$, so there is at least one element k with $\hat{D}_k(\mathbf{p}) < S_h$. For this k, $r_k(\mathbf{p}, \hat{\mathbf{D}}) > p_k$. ***Q.E.D.***

Now define a new function that is a truncated version of r_h: $\tilde{r}_h(\mathbf{p}, \hat{\mathbf{D}}) = \max\{r_h(\mathbf{p}, \hat{\mathbf{D}}), \underline{p}\}$. Clearly $\tilde{\mathbf{r}}(\mathbf{p}, \hat{\mathbf{D}})$ is a continuous function which maps $[\underline{p}, \overline{p}]^N$ into itself, so by Brouwer's fixed–point theorem, $\tilde{\mathbf{r}}(\mathbf{p}, \hat{\mathbf{D}})$ has a fixed point, $\mathbf{p}^*$. By the definition of $\underline{p}$ and $\overline{p}$, $\mathbf{p}^*$ cannot have a value at the lower bound, so $\mathbf{p}^*$ is in the interior of $[\underline{p}, \overline{p}]^N$. This implies that $\mathbf{p}^*$ is also a fixed point of the unrestricted function $\mathbf{r}(\mathbf{p}, \hat{\mathbf{D}})$, which establishes existence.

A well-known sufficient condition for uniqueness is $\sum_k |\partial r_h / \partial p_k| < 1$, which establishes that $\mathbf{r}$ is a contraction mapping. By the implicit function theorem, $\partial r_h / \partial p_k = -[\partial \hat{D}_h / \partial p_k] / [\partial \hat{D}_h / \partial p_h]$. From this $\sum_k |\partial r_h / \partial p_k| < 1$ if and only if the following dominant diagonal condition holds:

(A9) $$\sum_{k \neq (h,0)} \left| \partial \hat{D}_h / \partial p_k \right| < \left| \partial \hat{D}_h / \partial p_h \right|$$

To establish this condition, note that increasing all prices (including p_0) by the same amount will not change the demand for any house. Then (A9) follows from:

$$\sum_{k \neq h} \left| \partial \hat{D}_h / \partial p_k \right| - \left| \partial \hat{D}_h / \partial p_h \right| = 0 \quad \Rightarrow \sum_{k \neq (0,h)} \left| \partial \hat{D}_h / \partial p_k \right| = -\partial \hat{D}_h / \partial p_0 + \left| \partial \hat{D}_h / \partial p_h \right| < \left| \partial \hat{D}_h / \partial p_h \right|.$$

Q.E.D.

Proof of Lemma 1. Lemma 1 follows directly if we can show that the mapping that defines the fixed-point problem above is continuous in $\mathbf{x}$, as the *unique* fixed-point $\mathbf{p}^*$ of a mapping continuous in both $\mathbf{x}$ and $\mathbf{p}$ is also continuous in $\mathbf{x}$. The assumption that utility is continuous in x_h along with assumption about the continuous distribution of $\mathbf{e}$ implies that P_h^i is continuous in $\mathbf{x}$ for all i, which in turn implies that $\hat{D}_h$ is continuous in $\mathbf{x}$, which in turn implies that the element-by-element inverse defined in (A8) is continuous in $\mathbf{x}$. ***Q.E.D.***

Proof of Proposition 2. Conditional on any vector **g** and the primitives of the model {**Z, X, x**}, Proposition 1 implies that a unique set of housing prices clears the market and assumption (i) ensures that this vector of market-clearing prices is continuous in **g**. Assumptions (ii) and (iii) in turn imply that that equation (2.11), along with the definition of the function **g**, implicitly defines **g** and represents a continuous mapping of a closed interval into itself. The existence of fixed point of this mapping, $\mathbf{g}^*$, follows directly from Brouwer's fixed-point theorem. Any fixed point, $\mathbf{g}^*$, is associated with a unique vector of market clearing prices $\mathbf{p}^*$ and a unique set of choice probabilities $\{P_h^{i\,*}\}$ that together satisfy the conditions for a sorting equilibrium. Consequently, the existence of a fixed point, $\mathbf{g}^*$, implies the existence of a sorting equilibrium. ***Q.E.D.***

RESULTS APPENDIX

In order to judge whether our parameter estimates are reasonable, it is helpful to compare them to analogous hedonic price regressions. This Appendix carries out such a comparison, complementing the discussion in Section 5.

Hedonic price regressions arise as a direct restriction on our residential sorting model when there is no heterogeneity in household preferences for each house. See Bayer, Ferreira and McMillan (2003) for more details.[3] Equation (11), which describes mean preferences in the general case where preferences are heterogeneous, can be re-written:

$$\text{(A10)} \quad p_h + \tfrac{1}{\alpha_{0p}}\delta_h = \tfrac{\alpha_{0X}}{\alpha_{0p}} X_h + \tfrac{\alpha_{0Z}}{\alpha_{0p}} Z_h + \tfrac{1}{\alpha_{0p}}\xi_h$$

This bears more than a passing resemblance to a hedonic price regression. It makes clear that, in the presence of heterogeneous preferences, the mean indirect utility δ_h estimated in the first stage of the estimation procedure provides an adjustment to the hedonic price equation so that the price regression accurately returns mean preferences.

It is useful to spell out the significance of (A10). We can distinguish the willingness to pay of the *marginal* household, setting the equilibrium price of a given attribute, and that of the *mean* household. The equilibrium price function, approximated by a hedonic price regression, measures the marginal willingness to pay (MWTP) of the marginal household, and in the presence of heterogeneity, this may differ markedly from the MWTP of the mean household. The sorting model controls for which individual in the distribution of tastes sets the price of a given attribute given the supply of that attribute. This provides an adjustment that reflects the difference between this household's valuation and that of the mean household so that the adjusted hedonic price regression accurately reflects mean preferences.

The final two columns of Table 2 present the results from three hedonic price regressions analogous to those reported in the first three columns for the full sorting model. Comparing the hedonic price regressions to the mean MWTP estimates derived from the sorting model reveals that while the estimates related to housing characteristics, school quality, and crime remain similar in the hedonic price regression, those related to neighborhood sociodemographic composition and race in particular change dramatically. To explain the results, consider the estimated mean coefficient on percent black, which is -$234 in the full sorting model as opposed to only -$40 for the hedonic price regression. For simplicity, assume that neighborhoods are completely segregated, so that the equilibrium price of a black neighborhood is driven by the MWTP of the black household with the least MWTP for a black neighborhood (or, alternatively, the white household with the greatest MWTP). Here, the hedonic price regression returns the MWTP of the household on the *margin* between choosing a black versus white neighborhood, which in this case is substantially greater than the MWTP of the *mean* household, which is estimated in the more general sorting model. Put another way, a much lower differential in price between black and white neighborhoods is required to equilibrate the housing market than would be required to make the mean household indifferent between these neighborhoods.

---

[3] For a more careful discussion as to how the discrete choice model described here relates to continuous choice models commonly used in the hedonics literature (including Rosen (1974), Brown and Rosen (1982), Epple (1987), Bartik (1987), and Ekeland, Heckman, and Nesheim (2002), Bajari and Benkhard (2002)), see Bayer, McMillan, and Rueben (2003).