A UNIFIED FRAMEWORK FOR MEASURING PREFERENCES FOR SCHOOL QUALITY*

Patrick Bayer
Department of Economics
Yale University

Fernando Ferreira
Department of Economics
University of California, Berkeley

Robert McMillan
Department of Economics
University of Toronto

Preliminary and Incomplete
All Comments Welcome

October 2002

* We are grateful to Steve Berry, Sandra Black, Hanming Fang, Kim Rueben, Chris Taber, and Chris Timmins and participants at the 2002 Urban School Finance Workshop at the University of Illinois-Chicago for providing many valuable comments and suggestions. We also thank Pedro Cerdan and Jackie Chou for help in assembling the data. 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 the financial support for this project provided by the National Science Foundation under grant SES-0137289 and the Public Policy Institute of California.

1 INTRODUCTION

Much of the recent debate over choice-based education policies, such as vouchers and charter schools, has centered on the question of whether these policies will significantly increase the stratification of children across schools on the basis of ability, parental education, income, or race. The answer to this question depends critically on the nature of the fundamental factors driving the schooling and residential location decisions of the households in the relevant geographic area – the underlying variation in preferences for school quality, the extent to which these preferences are correlated with family characteristics such as education and income, and the relative importance of schooling in driving residential location decisions. The primary goal of this paper is to provide a comprehensive framework for recovering the distribution of preferences for school quality, a framework that brings together many of the best ideas that have been developed in the literature for handling a number of important endogeneity problems that complicate this problem. In the process of developing a single unified framework, we also seek to relate many of these previous approaches to one another, making clear the assumptions implicit in the simpler methodologies and the types of variation in the data needed for the identification of the more sophisticated ones.

The starting point for our analysis is the hedonic price regression framework of Black (1999), whose main contribution is a strategy for dealing with the likely correlation of school quality with unobserved housing and neighborhood characteristics. Using a sample of houses near a number of distinct school attendance zone boundaries, this strategy is to include boundary fixed effects in the regression, in essence comparing the prices of houses in otherwise similar neighborhoods, but that fall on opposite sides of a boundary determining where kids attend school. While Black (1999) suggests that this approach returns the mean marginal willingness to pay (MWTP) for school quality, well-known results in the literature on hedonic models dating back to Tinbergen (1956) and Rosen (1974) suggest that this is the case only in special circumstances. If, for example, all households had identical preferences (including over locations), prices would adjust in order to make each household indifferent among the full set of houses in the metropolitan area, thus ensuring that the equilibrium price of each house reflected the mean valuation of its attributes. In general, however, the equilibrium price of each house is a function of the full distribution of the household preferences driving the residential location decision as well as the geographic distribution of houses and neighborhoods (and their attributes) throughout the metropolitan region. Consequently, in order to properly estimate even just mean preferences for school quality it is necessary to explicitly consider the determination of equilibrium prices in the presence of a heterogeneous set of households and residences.

To this end, we develop an equilibrium model of sorting in an urban housing market. Building on the discrete choice framework developed in McFadden (1978) and extended to explicitly include unobserved choice characteristics by Berry, Levinsohn, and Pakes (1995), this model allows households to have preferences for a large number of housing and neighborhood attributes and for these preferences to vary with a wide variety of household characteristics including wealth, income, education, race, employment location, and family composition. We close the model with a market clearing condition that characterizes how the price of each house is determined in equilibrium. This equilibrium framework nests the hedonic price regression as a direct restriction (not surprisingly, when households have identical preferences) and, importantly, it remains possible to incorporate the Black boundary fixed effects approach within the context of this broader model.

The general methodology that we develop requires more data than were used in Black's analysis and consequently, we turn to newly available restricted-access Census data. These data provide detailed information on a 1-in-7 sample of households, including each household member's race, education, income, age, immigration status, employment status, and job location. Importantly, unlike publicly available Census data, which match each household with a PUMA (a Census area of at least 100,000 residents), these provide a household's residential and employment locations at the level of a Census block (a Census area with approximately 100 residents), allowing us to characterize each household's actual location much more accurately than has been possible in past studies that have used Census micro-data. Using these new Census data as a centerpiece, we have assembled an extensive data set characterizing the housing market in the San Francisco Bay Area, combining housing and neighborhood sociodemographic data drawn from the Census with neighborhood-level data on schools, air quality, climate, crime, topography, geology, land use, and urban density.

One of the central goals of this paper is to provide a clear characterization of the sources of variation in the data that determine the model's parameters under alternative specifications. We begin by discussing the sources of variation in the data that distinguish the broader equilibrium model of sorting (which includes observed and unobserved heterogeneity in preferences) relative to the hedonic price regression. Our analysis uses cross-sectional data from a single, large metropolitan area market and, consequently, the identification of the unobserved heterogeneity in preferences (random coefficients) is completely driven by geographic variation. In particular, we condition on the geographic distribution of employment, thereby introducing variation in the way that households situated in various parts of the metropolitan area view the set of houses available in the market. In addition to ensuring reasonable spatial substitution patterns,

by observing the location decisions made by observably identical households when facing different choice sets, we are able to learn about substitution patterns more generally, thereby distinguishing the underlying heterogeneity in tastes.[1]

In the context of this discussion, we draw a distinction between the use of *choice variation* (differences between the alternatives that are chosen vs. not chosen by households of a particular type) versus *within-type price variation* (the variation in prices and attributes among the set of houses chosen by households of a particular type) in identifying the model's parameters. We demonstrate that when the estimation of preferences is based solely on within-type price variation a classic sample selection bias arises, as the unobserved quality level of the houses chosen by a given type of household is likely to be correlated with the observed attributes of these houses. Among the set of houses chosen by high-income households, for example, relatively small houses or houses with poor neighborhood amenities are likely to have high levels of unobserved house or neighborhood quality, leading to a downward bias in the estimation of the preferences of high-income households for the observed amenities.

We also demonstrate that when preference estimates are derived using either only within-type price variation or a standard discrete choice approach, which uses both forms of variation, another form of selection bias arises if school quality is correlated with unobserved housing or neighborhood attributes that are valued more strongly by one type of household versus another. If, for example, high-income households have relatively strong preferences for both good schools and homes with a view of the San Francisco Bay, these analyses would overstate the relative preferences of high-income households for school quality. We provide a solution to this selection problem that essentially uses choice variation across neighborhoods to absorb out type-specific differences in the value of unobserved neighborhood attributes, leaving differences in the choice of houses within neighborhoods and the within-type price variation as the basis for estimating preferences for school quality. In this way, the final specification that we develop provides consistent estimates of the distribution of preferences for school quality even in the presence of a wide variety of likely endogeneity problems.

[results not yet disclosed for inclusion in paper]

The final estimates of our equilibrium model of sorting serve a number of potentially interesting purposes. Most obviously, they provide a clear characterization of the mean, observable heterogeneity, and the unobserved variation in preferences for school quality. Equally as important, the estimated model provides a well-defined characterization of the relative

[1] Of course, conditioning on the geographic distribution of employment with the metropolitan area is not a costless assumption. We discuss the likely impact of this assumption on the results in Section 4 below.

importance of schooling versus other housing, neighborhood, and geographic factors in driving the location decisions of the heterogeneous households of a major metropolitan area. This combination is extremely powerful for conducting economic and policy research involving the interplay of household mobility/stratification and schools. This model makes it possible, for example, to calculate the elasticity of neighborhood house prices and rents as well as the sociodemographic composition of the local neighborhood and school with respect to school quality for each school in the metropolitan region. These 'demand' elasticities provide a series of measure of the competitiveness of a school's local environment and can be used to explore many aspects of the interplay between household mobility and school competition.[2] Moreover, the sorting model provides a way of calculating the strength of preferences for school quality (on the basis of both observed and unobserved characteristics) among the households that select into a particular school. This permits the researcher to control directly for the non-random sorting of households across schools and school districts which leads to a form of selection bias (often referred to as Tiebout bias in the local public finance literature) in the estimation of education production functions, voting models, or other models that condition implicitly on the set of households in a particular school or jurisdiction.[3]

2 AN EQUILIBRIUM MODEL OF THE URBAN HOUSING MARKET

This section of the paper develops our equilibrium model of the urban housing market. The model consists of two key elements: the household residential location decision problem and a market clearing condition that that characterizes how the price of each house is determined in equilibrium. We begin with a specification in which all households have identical preferences for houses and neighborhoods up to an idiosyncratic term in the utility function. In this case, our equilibrium framework reduces to a standard hedonic price regression, which accurately returns mean preferences in this case. We then extend the model to the case where households have heterogeneous tastes for housing and neighborhood attributes as well as locations.

2.1 A Model with Homogeneous Preferences

We model the residential location decision of each household as a discrete choice of a single residence. The utility function specification is based on the random utility model developed in McFadden (1978) and the specification of Berry, Levinsohn, and Pakes (1995), which includes choice-specific unobservable characteristics. In the model, each household

[2] See Bayer, McMillan, and Rueben (2002b).

chooses its residence h to maximize its utility, which depends on the observable and unobservable characteristics of its choice. Let X_h represent the observable characteristics of house h other than price that vary with the household's housing choice and let p_h denote its price. The observable characteristics of a housing choice include 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., sociodemographic composition, school, crime, topography, and air quality).

When households have homogeneous preferences up to an idiosyncratic term, household i's optimization problem is given by:

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

where ξ_h is the unobserved quality of each house, including any unobserved quality associated with its neighborhood, and ε^i_h is an idiosyncratic error term that captures unobserved variation in household i's preference for a particular housing choice.

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

$$(2) \qquad V_h^i > V_k^i \quad \Rightarrow \quad W_h^i + \varepsilon_h^i > W_k^i + \varepsilon_k^i \quad \Rightarrow \quad \varepsilon_h^i - \varepsilon_k^i > W_k^i - W_h^i \qquad \forall \quad k \neq h$$

where W^i_h includes all of the non-idiosyncratic components of the utility function V^i_h. As the inequalities depicted 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 choices. In this way, the probability P^i_h that household i chooses house h can be written as a function of the full vectors of house characteristics (both observed and unobserved) and prices $\{\mathbf{X}, \mathbf{p}, \boldsymbol{\xi}\}$:

$$(3) \qquad P_h^i = f_h(\mathbf{X}, \mathbf{p}, \boldsymbol{\xi})$$

Equilibrium[4,5]

---

[3] In Bayer, McMillan, and Rueben (2002c), we exploit this feature to properly estimate the role of family characteristics in the production of achievement.

[4] For more details on the assumptions underlying the equilibrium concept used in this analysis see Bayer, McMillan, and Rueben (2002), which extends the equilibrium analysis to include social interactions. For

When the set of draws $\{e^i_h\}$ for each household observed in the data is interpreted as idiosyncratic heterogeneity in preferences for each house, working with choice probabilities is equivalent to assuming that each household that we observe in our sample represents a continuum of households with the same observable characteristics. The choice probabilities shown in equation (3) depict the distribution of location decisions that would result for a continuum of households with a given set of observed characteristics as each household responds to its particular idiosyncratic preferences. Let the measure of the continuum of households be m. This assumption concerning the distribution of households requires a similar assumption about the set of housing choices observed in the sample. In order to make the model coherent, therefore, we also assume that each house observed in the sample represents a continuum of identical houses, and that this continuum also has measure m.

Aggregating the probabilities in equation (4) over all households yields the predicted number of households that choose each house h, $\hat{N}_h$:

$$\hat{N}_h = m \bullet \sum_i P^i_h \tag{4}$$

where again m represents the measure of the continuum of households with the same observable characteristics as household i. In order for the housing market to clear, the number of households choosing each house h must equal the measure of the continuum of houses that each observed house represents:[6]

$$\hat{N}_h = m, \quad \forall h \quad \Rightarrow \quad \sum_i P^i_h = 1, \quad \forall h \tag{5}$$

When households have homogeneous preferences, this market clearing condition implies that prices adjust so that the mean indirect utility that each house provides is identical in equilibrium and, consequently:

---

clarity of exposition, we ignore such interactions in presenting the model and equilibrium properties in this paper.

[5] The equilibrium concept developed here treats the supply of housing as fixed. This is done for expositional simplicity as a more generic housing supply function could certainly be incorporated in the analysis.

[6] Note that the measure m drops out of the market-clearing condition depicted in equation (8) and, consequently, simply serves as a rhetorical device for understanding the use of the continuous choice

$$(6) \qquad \alpha_{0X} X_h - \alpha_{0p} p_h + \xi_h = K \qquad \Rightarrow \qquad p_h = {}^{\alpha_{0X}}\!/_{\alpha_{0p}} X_h + {}^{1}\!/_{\alpha_{0p}} \xi_h$$

Equation (6) is the standard hedonic price regression.

2.2 A Model with Heterogeneous Preferences

We now introduce a utility specification that allows households to have heterogeneous preferences for house and neighborhood attributes as well as locations.[7] As before, the residential location decision of each household is modeled as a discrete choice of a single residence. In this case, however, household i's optimization problem is given by:

$$(7) \qquad \underset{(h)}{Max} \quad V_h^i = \alpha_X^i X_h - \alpha_D^i D_h^i - \alpha_p^i p_h + \xi_h + \varepsilon_h^i$$

where the $\alpha^i{}_D D^i{}_h$ term in the utility function captures the disutility of commuting – the negative impact of the distance between household i's workplace and house h – and 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. We also assume that each working household is initially endowed with a primary employment location, l^i. We treat employment status and employment location as exogenous variables throughout this paper. Each parameter associated with housing and neighborhood characteristics, distance to work, and price, α^i_j, for $j \in \{X, D, p\}$, is allowed to vary with a household's own characteristics,

$$(8) \qquad \alpha_j^i = \alpha_{0j} + \sum_{r=1}^{R} \alpha_{rj} Z_r^i + \eta_j^i ,$$

In this way, equation (8) describes household i's preference for choice characteristic j. The first term captures the taste for the choice characteristic that is common to all households and the other terms capture observable variation in the valuation of these choice characteristics across

---

probabilities shown in equation (6) in defining equilibrium rather than the actual discrete choices of the individuals observed in the data.

[7] This section summarizes the key aspects of the equilibrium model of sorting in an urban housing market developed in Bayer, McMillan, and Rueben (2002).

households with different socioeconomic characteristics. This random coefficients specification allows for great variation in preferences across different types of household.[8]

In this case, the probability P^i_h that household i chooses house h can be written as a function of the full vectors of house characteristics (both observed and unobserved) and prices $\{\mathbf{X}, \mathbf{p}, \mathbf{x}\}$:

$$(9) \qquad P^i_h = f_h(Z^i, \boldsymbol{\nu}^i, \mathbf{X}, \mathbf{p}, \mathbf{x})$$

as well as the household's own characteristics Z^i, ν^i.[9] Given the same market clearing conditions as above, it is a straightforward extension of the central proof in Berry (1994) to show that a unique vector of housing prices clears the market.[10] Writing this market-clearing vector of prices as $\mathbf{p}^*(\mathbf{v}, \mathbf{Z}, \mathbf{X}, \mathbf{x})$, the probability that household i chooses house h can be written:

$$(10) \qquad P^i_h = f_h\left(Z^i, \boldsymbol{\nu}^i, \mathbf{X}, \mathbf{p}^*(\mathbf{Z}, \mathbf{X}, \hat{\mathbf{i}}), \hat{\mathbf{i}}\right)$$

where the notation $\mathbf{p}^*(\mathbf{v}, \mathbf{Z}, \mathbf{X}, \mathbf{x})$ indicates that the set of market-clearing prices is generally a function of the full matrices of the household $\{\mathbf{v}, \mathbf{Z}\}$ and house and neighborhood characteristics $\{\mathbf{X}, \mathbf{x}\}$ that are treated as the primitives of the sorting model. Consequently, while the hedonic price function reflects mean household preferences when preferences are homogeneous, as just described, equilibrium prices are generally a function of the set of available options and the distribution of tastes within the population. An equilibrium is defined as the set of choice probabilities in equation (10) along with the vector of market clearing prices, $\mathbf{p}^*$. Since a unique set of prices clears the housing market, the sorting equilibrium will also be unique when the model does not include any social interactions.[11]

[8] For a more detailed discussion of the stochastic elements in this specification see Bayer, McMillan, and Rueben (2002).

[9] For simplicity of exposition, we have included the household's employment location in Z^i and the location of the house in X_h. Note also that the h subscript on the function f simply indicates that we are solving for the probability that household i chooses house h not that the form of the function itself varies with h.

[10] See Bayer, McMillan, and Rueben (2002) for related proofs.

[11] Again for a more detailed discussion of the equilibrium definition and properties with social interactions see Bayer, McMillan, and Rueben (2002).

Estimation

We now present a summary of the procedure that we use to estimate the model with heterogeneous preferences.[12] As discussed above, the homogenous preference case simply reduces to a standard hedonic price regression. We begin by introducing some notation that simplifies the exposition. The terms of the utility function specified in equations (1)-(2) can be divided into a *choice-specific constant,* δ_h, an *interaction component,* $\mu^i{}_h$, which includes any parts of the utility function that interact household and choice characteristics, and the *idiosyncratic error term,* $\varepsilon^i{}_h$. Thus the utility function can be rewritten as:

$$(11) \qquad V^i{}_h = \delta_h + \mu^i{}_h + \varepsilon^i{}_h.$$

where:

$$(12) \qquad \delta_h = \alpha_{0X} X_h - \alpha_{0p} p_h + \xi_h$$

$$(13) \qquad \mu_h^i = \left(\sum_{k=1}^{K} \alpha_{kX} Z_k^i \right) X_h - \left(\alpha_{0D} + \sum_{k=1}^{K} \alpha_{kD} Z_k^i \right) D_h^i - \left(\sum_{k=1}^{K} \alpha_{kp} Z_k^i \right) p_h$$

In these equations, k indexes household characteristics. The choice-specific constant δ_h captures the portion of the utility provided by house h that is common to all households. When the household characteristics included in the model are constructed to have mean zero, δ_h is the mean indirect utility provided by house h. The unobservable component of δ_h, ξ_h, captures the portion of unobserved preferences for house h that is correlated across households, while $\varepsilon_h{}^i$ represents unobserved idiosyncratic preferences over and above this shared component.[13] Denoting the full set of parameters θ, we subdivide these into two sets in later discussion: the set of interaction parameters in $\mu^i{}_h$, θ_μ, and the set of parameters in δ_h, θ_δ.

The estimation of the model begins by maximizing the probability that each household chooses its observed location. In particular, for any combination of interaction parameters and choice-specific constants, δ_h, the model predicts the probability that each household i chooses house h:

[12] Complete details of the estimation procedure including methods for simplifying the computation and the asymptotic properties of the estimator can be found in Bayer, McMillan, and Rueben (2002).

$$(14) \qquad P_h^i = \exp(\delta_h + \hat{\mu}_h^i) \Big/ \sum_k \exp(\delta_k + \hat{\mu}_k^i)$$

Maximizing the probability that each household makes its correct housing choice, conditioning on the full set of observed household characteristics Z^i and choice characteristics $\{X_h, p_h, \overline{Z}_h\}$, gives rise to the following log-likelihood function:

$$(15) \qquad \ell = \sum_i \sum_h I_h^i \ln(P_h^i)$$

where I^i_h is an indicator variable that equals 1 if household i chooses house h in the data and 0 otherwise. The first step of the estimation procedure consists of searching over the interaction parameters and vector of choice-specific constants to maximize ℓ, returning estimates of the interaction parameters $\hat{\pi}_m$ and the vector of choice-specific constants $\hat{\delta}$.

Notice, however, that the set of observed residential choices provides no information that distinguishes the elements of the choice-specific constant $\boldsymbol{\delta}$. Consequently, it is necessary to bring additional econometric information to bear on the problem. Given the estimate of $\boldsymbol{\delta}$ obtained from fitting the observed individual location decisions, equation (12) is simply a regression equation. The most obvious approach to identifying the parameters of this equation involves forming moments based on covariance restrictions between the observed choice characteristics and x_h. It is clear, however, that forming covariance restrictions between x_h and p_h is not consistent with the logic of the choice model, as any increase in the unobserved quality of a house typically raises demand for a house and in turn its equilibrium price. The second step of the estimation procedure, therefore, uses $\hat{\delta}$ along with an appropriate instrument for price to estimate equation (12) via instrumental variables. We discuss the formation of appropriate instruments in the next section of the paper.

3 ENDOGENEITY ISSUES

A number of important endogeneity issues arise in the estimation of preferences for school quality. In this section of the paper, we discuss three such issues in detail: (i) the self-selection across locations when households have heterogeneous tastes; (ii) the correlation of

[13] Another way to describe ξ_h is that it captures the shared portion of the quality of house h (including the

school quality with unobserved housing and neighborhood quality; and (iii) the correlation of school quality with unobserved neighborhood characteristics valued more heavily by some households versus others. For each of these issues, we describe the endogeneity issue, present a strategy for identifying the model, and discuss the sources of variation in the data that implicitly tie down the preference parameters.

3.1 The Hedonic Identification Problem with Unobservable Choice Characteristics

A central identification issue in the paper concerns the endogeneity of housing prices, p_h. As we discussed briefly in describing the estimation procedure, the identification of the choice-specific constant regression (equation (12)) requires a set of instruments that are correlated with p_h but not with unobserved housing/neighborhood quality, x_h. In developing appropriate instruments for p_h, we provide a solution to the classic problem of distinguishing preferences from the equilibrium (hedonic) price function.

The instruments that we construct rise naturally out of the sorting model when households value only the characteristics of their chosen house and the features of the surrounding neighborhood, as long as the geographic extent of this 'neighborhood' is reasonably small relative to the full metropolitan area. In this case, the exogenous attributes of houses and neighborhoods which are positioned just beyond the region that households value directly make ideal instruments for housing prices. In particular, for each house h, we form instruments that characterize the housing stock and general land usage patterns (percent residential, commercial, industrial, etc.) beyond a threshold distance from house h. At the same time, we are careful to include variables that describe the housing stock and land usage within this threshold distance directly in the utility function. In developing this set of instruments, we exploit an inherent feature of the sorting process – that the overall demand for houses in a particular neighborhood is affected by not only the features of the neighborhood itself, but also by the way these features relate to the broader landscape of houses and neighborhoods in the region. In this way, we assume that the exogenous attributes of nearby but not immediately proximate neighborhoods influence the equilibrium in the housing market, thereby affecting prices, but have no direct effect on utility.

---

quality of its neighborhood) that is observed by the households in the data but not the econometrician.

To demonstrate the general logic of this first category of instruments, the first column of Table 1 reports the results of a first-stage regression that relates the price of house *h* to a series of attributes of the house and neighborhood as well as a set of variables that characterize the housing stock and land usage in concentric rings surrounding house *h*. Here, we make use of the highly disaggregated geographic information in the dataset and for each Census block construct variables that characterize the housing stock and land usage in rings with radii of 0-to-1, 1-to-3, and 3-to-5 miles.[14] The results of the regression shown in the first column of Table 1 highlight the logic of our instrumental variables strategy. Consider, for example, the impact of industrial land usage on price. While industrial land usage within a mile of a house has a significantly negative impact on prices, this effect drops considerably in the 1-to-3 mile ring, and then turns *positive* in the 3-to-5 mile ring. Interpreted in the context of the equilibrium model, this sign change suggests the presence of market forces at work – that is, the presence of more industrial land 3-5 miles away reduces the supply of and lowers the quality of houses in relatively close geographic proximity to the neighborhood in question, thereby increasing the equilibrium price of houses in that neighborhood.[15]

Comparing the estimated parameters for the land use and neighborhood housing stock variables within a mile to the estimated parameters for these same variables 3-to-5 away, statistically significant sign changes are occur for the all of other land use measures and four of the six variables that characterize the housing stock. In the two cases in which the coefficient on the housing variable does not change sign, the estimated coefficient declines by an order of magnitude or more. In estimating the model, then, we select 3 miles as our threshold distance, using variables that characterize the land use and housing stock in the 3-to-5 mile ring as instruments for price and neighborhood sociodemographic characteristics, including variables that characterize the land use and housing stock within 3 miles of the house directly in the utility function. These instrumental variables are jointly significant in first stage regressions, as the F-statistics reported in Table 1 *(F = 181.98)* makes clear.

[14] Summary statistics for these variables are presented in Appendix Table 1.

The Classic Hedonic Identification Problem

While typically framed in a slightly different equilibrium framework, the identification of our model (ignoring the additional burdens of identifying the components related to neighborhood sociodemographics) is essentially the same as the identification problem that underlies a hedonic model with unobserved choice characteristics – i.e., that of distinguishing the equilibrium price gradient from the marginal utility of households.[16] Solutions to the problem of identifying hedonic models with unobserved choice characteristics have only recently begun to appear in the literature and, consequently, we present a short discussion of the issues related to the identification of our model in order to facilitate comparison of the assumptions underlying identification.[17]

Following Rosen (1974), the classic strategy for estimating hedonic models involves estimating an equilibrium (hedonic) price function in a first-stage regression and using the gradient of this price function as the dependent variable in the household's first-order conditions in a second stage. This second stage of this classic estimation strategy returns both the mean and variation in preferences. Our estimation strategy begins instead by estimating the heterogeneity in preferences along with a vector of mean utilities d in the first stage, identifying mean preferences in the second stage - the choice-specific constant regression.[18] In this way, the identification of our model is tied to these two stages of estimation.

Consider first the identification of the second stage regression, which uses the vector of choice-specific constants d estimated in the first stage. Ignoring the endogeneity of neighborhood

---

[15] Note that this regression controls directly for a series of employment access variables that control for the fact that increased industrial or commercial land usage 3-5 miles out may simply pick up better employment access.

[16] The hedonic literature typically uses a continuous choice framework. Important contributions include Rosen (1974), Brown and Rosen (1982), Epple (1987), Bartik (1987), and Ekeland, Heckman, and Nesheim (2002). While the underlying assumptions implicit in standard discrete choice estimation differ from the standard estimation of a hedonic model (which generally assumes that the observed decision satisfies a first order condition w.r.t. each choice attribute) the fundamental nature of the identification problem, distinguishing the equilibrium price function from preferences, is the identical in both approaches.

[17] A new working paper by Bajari and Benkhard (2002) presents alternative methods for identifying hedonic models of demand with unobserved product characteristics. It should also be noted that the general nature of the portion of our identification strategy related to the hedonic identification problem is derived from a long line of research in IO starting with the work of Bresnahan (1981, 1987) and Berry, Levinsohn, and Pakes (1995).

[18] Throughout this discussion, we assume that the mean has been subtracted from all included household characteristics so that the choice-specific constant regression returns mean preferences.

sociodemographic characteristics for the purposes of this discussion, equation (4.2) contains two endogenous variables: housing price, p_h, and mean utility, d_h. Recall from Section 2, that when households are identical except for their idiosyncratic locational preferences, the market clearing condition implies that the mean utility of each house must be identical in equilibrium. In this case, then, one of the endogenous variables is eliminated and mean preferences can be recovered by moving price to the other side of equation (12) and simply regressing price on characteristics. In other words, without household heterogeneity in preferences or geography, and with the assumption that housing supply is fixed, the equilibrium price function simply reflects mean preferences. From the point of view of the market clearing conditions, the logic is simple: any increase in the unobserved quality of a house, x_h, is immediately offset by an increase in p_h, leaving mean utility and individual location decisions unchanged.

When heterogeneity in preferences or geography is allowed for, an increase in the unobserved quality of a choice alters the decisions that are made in equilibrium. If, as we find in our analysis, unobserved quality is a normal good, an increase in the unobserved quality of a house/neighborhood generally increases the income of the households that locate there in equilibrium. In this context, then, the presence of d_h in equation (12) provides the appropriate adjustment to the hedonic price equation to return mean household preferences:

$$(17) \qquad p_h + \frac{1}{\alpha_{0p}} d_h = \frac{\alpha_{0X}}{\alpha_{0p}} X_h + \frac{1}{\alpha_{0p}} x_h$$

This mean-utility adjustment generally depends in equilibrium on the distribution of households and their tastes, the geographic distribution of employment, and the geographic distribution of houses and neighborhoods and their characteristics. It is in providing a valid instrument for p_h, then, that we identify the second stage of our two-stage estimation procedure.

The identification of heterogeneous preferences and the vector of choice-specific constants in the first stage of the estimation procedure for the random coefficients model presented in Section 2 guaranteed is driven by our assumption concerning the geography of employment locations. From a practical point of view, the identification of random coefficients derives from the same source that generates variation in our proposed instruments for price –

namely the geography of the location decision. In particular, in our framework, the geographic distribution of employment within the Bay Area gives each household a distinct geographic bliss point, leading to tremendous variation in a household's perception of the choice set depending on the location of its workplace(s) within the metropolitan area. It is this variation that makes choices closer in geographic space closer substitutes for one another, thereby giving our instruments, which describe the local housing and labor market conditions for each residential choice, empirical content. This variation also allows us to learn about substitution patterns by observing households with identical observable characteristics choosing from different sets of choices in different parts of the metropolitan area. By estimating substitution patterns, one is able to reject the substitution patterns predicted by the multinomial logit in the data and thereby distinguish more flexible forms for the stochastic structure of the preferences. In this way, the geographic variation in preferences introduced by conditioning on the distribution of employment within the Bay Area is essential for the identification of the model. One final point is worth emphasizing: while we estimate the model using data from a single large metropolitan area, the problem is not what is usually referred to as identification using data from a single market, which typically implies that the researcher does not observe variation in the choice set.

Optimal Instruments

In forming an instrument for housing prices, we exploit the fact that the demand for houses in any particular neighborhood is determined not only by the features of that neighborhood, but also by how these features relate to the broader landscape of housing/neighborhood choices. The instruments developed above along with the first stage regression for housing prices shown in the first column of Table 1 are important precisely because they demonstrate the logic of this identification strategy. In general, however, because we would like to instrument for a large number of choice characteristics, the precision of the estimation is improved significantly with the use of a parsimonious set of instruments that approximate the optimal instruments for price in our econometric framework. This sub-section of the paper characterizes the optimal instruments and develops computable instruments that approximate them.

The optimal instruments for p_h in the choice-specific constant regression (equation (12)) are given by:

$$(18) \qquad E\left(\partial x_h / \partial a_{0p}\right) = E(p_h \mid \Omega)$$

that is, the expected value of p_h conditional on the information set Ω, which contains the full distribution of *exogenous* choice (X_h) and individual characteristics (Z^i). Notice that these instruments implicitly incorporate the impact of the full distribution of the set of choices in exogenous characteristic space as well as information on the full distribution of observable household characteristics into a single instrument for each endogenous variable.

Because the equilibrium in the sorting model is not generically unique, however, this expectation is not well-defined. In particular, the calculation of this expectation requires computing the equilibrium for the full distribution of possible parameter values and the vector of unobserved choice characteristics, $\mathbf{x}$. Since some parameter values give rise to multiple equilibria, the expectation cannot be calculated without some way of determining how an equilibrium is chosen in these cases. For this reason, we use a well-defined instrument that maintains much of the inherent logic of this optimal instrument while being straightforward to compute. This 'quasi-' optimal instrument is based on the predicted vector of market-clearing prices calculated for an initial consistent estimate of the parameter values with the vector of unobserved characteristics $\boldsymbol{x}$ set identically equal to $\mathbf{0}$. This condition corresponds to using the prediction at the mean instead of the expected value.

The calculation of these instruments requires a consistent estimate of the model's parameters. Notice, however, that we can use the standard instruments developed above to provide an initial consistent estimate of these. Operationally, then, the estimation proceeds as follows:

1. While controlling for characteristics of housing stock and land usage within 3 miles, use instruments that characterize the housing stock and land usage 3-5 miles away to estimate choice-specific constant regression.

2. Using the resulting consistent parameter estimates, setting $x_h=0$ for all h, calculate the vector of housing prices that clears the market, $\hat{p}^*(X_h, Z^i)$. The notation here is intended to indicate the predicted vector of market clearing prices conditional solely on the observable, exogenous characteristics of households and houses/neighborhoods.

3. Using $\hat{p}^*$ as an instrument for p, estimate the choice-specific constant regression.

Like the optimal instrument, the instrument that we propose provides a measure of the way that the full landscape of possible choices affects the demand for each house/neighborhood. In essence, these instruments extract additional information from Ω than is contained in the vectors of choice characteristics **X**, which are already used directly in estimating equation (12). Moreover, the single 'quasi-' optimal instrument combines this information in a concise manner that is consistent with the logic of the sorting model. The final two columns of Table 1 show first-stage price regressions that include the 'quasi-' optimal instruments, both with and without the standard set of variables shown in the first column of the table. Even conditional on the full set of standard instruments, the optimal instruments have strong predictive power (the t-statistic on the optimal price instrument is greater than 100) and tests for the joint significance of the full set of instruments and the optimal instruments alone strongly reject the null in both specifications.

3.2 The Correlation of School Quality with Unobserved Neighborhood Quality

Even in the context of the hedonic price regression, it is easy to see that an endogeneity problem is likely to arise, as the quality of local schools is likely to be positively correlated with unobserved housing and neighborhood quality. In this way, simply estimating the model via OLS is likely to overstate household's willingness-to-pay for school quality – misattributing preferences for unobserved house and neighborhood quality as tastes for schools quality. To address this issue, we follow the identification strategy developed in Black (1999). Using a sample of houses near school attendance zone boundaries, Black estimates a hedonic price regression that includes boundary fixed effects. By including boundary fixed effects, this strategy essentially compares the prices of houses in otherwise similar neighborhoods, but that fall on opposite sides of a boundary determining where kids will attend school. Any differences in prices not associated with housing characteristics are then be interpreted as the marginal willingness-to-pay for school quality.

In order to incorporate boundary fixed effects in the estimation, we assign each house to a region r. When a house is close to a boundary between two school districts, it will fall into a *boundary region* and when a house is more centrally located within a school district, it will fall

into a *central region*. Letting y_r be a region fixed effect for the region r to which house h belongs, we can re-write the utility function shown in equation (1) as:

$$(18) \qquad \underset{(h)}{Max} \quad V_h^i = \alpha_X^i X_h - \alpha_D^i D_h^i - \alpha_p^i p_h + \gamma_r + \xi_h + \varepsilon_h^i$$

Having accounted for a region fixed effect, ξ_h now represents the unobserved quality associated with the particular housing unit h within region r.

[Add additional discussion of assumptions implicit in boundary fixed effects approach – we can, for example, control for neighborhood and school sociodemographic characteristics on opposite sides of the border.]

A final issue for the analysis concerns the treatment of houses not near a school district boundary. In essence, while we seek to use only the variation in the data at the boundaries to estimate preferences for school quality, the logic of the choice model developed in Section 2 requires the use of all houses in the choice set. Notice, however, that given the specification of equation (16), equations (11)-(12) become

$$(19) \qquad \delta_h = \alpha_{0X} X_h - \alpha_{0p} p_h + \gamma_r + \xi_h$$

$$(20) \qquad \mu_h^i = \left(\sum_{k=1}^{K} \alpha_{kX} Z_k^i \right) X_h - \left(\alpha_{0D} + \sum_{k=1}^{K} \alpha_{kD} Z_k^i \right) D_h^i - \left(\sum_{k=1}^{K} \alpha_{kp} Z_k^i \right) p_h$$

That is, the boundary fixed effect appears only in the choice-specific constant regression. Thus, the first stage of the estimation procedure remains unchanged, returning estimates of the interaction parameters and the choice-specific constants. In the second stage of the estimation procedure, (i.e., the estimation of equation (19)), we use only the sample of houses in boundary versus central regions. In this way, the estimation of the interaction (heterogeneity) parameters in the utility function shown in equation (20) is based on the full sample of houses, while the estimation of the mean preference parameters (those in equation (19)) is based only on across-boundary variation in prices. And, when preferences are restricted to be homogeneous, our equilibrium model reduces to the analysis in Black (1999).

3.3 Heterogeneity in Tastes for Unobservable Neighborhood Attributes

The baseline model with heterogeneous preferences developed in Section 2 accounts for selection associated with the interaction of (observable and unobservable) household

characteristics and endowments with observable housing and neighborhood characteristics. The third endogeneity problem that we attempt to address concerns the possible correlation of school quality with unobservable housing and neighborhood attributes valued more strongly by some households versus others. If, for example, high-income households have relatively strong preferences for both good schools and homes with a view of the San Francisco Bay, our baseline analysis would overstate the relative preferences of high-income households for school quality.

In terms of our notation, the baseline specification shown in equation (1)-(2) restricts the valuation of unobserved housing and neighborhood quality, x_h, to be identical across all households. The following utility specification allows this valuation to vary by observable household types t, where household type might indicate households with different levels of education, different races, or different levels of income:[19]

$$(21) \qquad \underset{(h)}{Max} \quad V_h^{i,t} = \alpha_X^{i,t} X_h - \alpha_D^{i,t} D_h^i - \alpha_p^{i,t} p_h + \xi_h^t + \varepsilon_h^{i,t}$$

where

$$(22) \qquad \alpha_j^{i,t} = \alpha_{0j}^t + \sum_{r=1}^{R} \alpha_{rj}^t Z_r^i$$

We refer to this specification as a 'type-specific' discrete choice model. This specification also allows each taste parameter to vary by type, which both permits for a tremendous amount of heterogeneity in preferences and, as we discuss below, greatly simplifies the computation. Notice also that this specification is much broader than simply allowing each household type to have different preferences for unobserved housing and neighborhood characteristics. In the specification of (21)-(22), households of one type might place a positive value on an unobservable feature while households of another type might place a negative value.[20]

Given the specification shown in equations (21)-(22), the analogous notation to that of equations (11)-(12) becomes:

$$(23) \qquad \delta_h^t = \alpha_{0X} X_h^t - \alpha_{0p} p_h^t + \xi_h^t$$

$$(24) \qquad \mu_h^{i,t} = \left(\sum_{k=1}^{K} \alpha_{rX}^t Z_k^i \right) X_h - \left(\alpha_{0D}^t + \sum_{k=1}^{K} \alpha_{rD}^t Z_k^i \right) D_h^i - \left(\sum_{k=1}^{K} \alpha_{rp}^t Z_k^i \right) p_h$$

[19] Interactions of unobservable household and unobservable housing characteristics are of course already accounted for in the idiosyncratic error term.

[20] The placement of an IKEA in East Palo Alto, for example, has been hotly contested with some residents clearly viewing it as an amenity and others as a disamenity.

Notice that in this case, for each type t, the first stage of the estimation procedure returns d^t_h=-¥ for houses not chosen by households of type t. In other words, the model can perfectly predict the fact that households of type t do not choose certain houses simply by forcing type t's choice-specific constant for that house to negative infinity.

To see the issues involved with estimating this specification, it is again helpful to consider a restriction. In particular, consider a utility specification in which preference parameters vary only by type:

$$(25) \qquad \underset{(h)}{Max} \quad V_h^{i,t} = \alpha_X^t X_h - \alpha_p^t p_h + \xi_h^t + \varepsilon_h^{i,t}$$

The elimination of within-type heterogeneity implies that all houses chosen by type t must yield the same indirect utility. At the same time, the model can again perfectly predict the fact that households of type t do not choose certain houses simply by forcing type t's choice-specific constant for that house to negative infinity. In this way, for each type t, the first stage of the estimation procedure returns d^t_h=K, a finite constant, for all houses chosen by type t and d^t_h=-¥ for houses not chosen by households of type t.

The question then is how to proceed in the second stage regression. For a given type t, consider first the idea of dropping all observations with d^t_h=-¥. In this case, the model reduces to a hedonic price regression using only houses chosen by households of type t:

$$(26) \qquad p_h = \alpha_X^t / \alpha_p^t \, X_h + 1/\alpha_p^t \, \xi_h^t \quad \forall \quad h \quad chosen by type \quad t$$

For obvious regions we refer to equation (26) as a 'type-specific' hedonic price regression By running such a hedonic price regression for each type, we estimate the heterogeneity in preferences across types. In essence, this procedure uses only the price variation in houses chosen by type t to estimate preferences of households of type t. In the context of the boundary fixed effects, this procedure returns estimates of preferences of type t for differences in school quality by comparing what a household of type t is willing to pay for an equivalent house on opposite sides of a school district boundary.

While such a type-specific across-boundary comparison is intuitively appealing as a way of estimating the heterogeneity in preferences, the procedure just outlined (in which we drop all observations with d^t_h=-¥) is clearly subject to a form of sample selection bias, as we have

selected on the dependent variable in equation (22). In this way when the estimation of preferences is based solely on within-type price variation a classic sample selection bias arises, as the unobserved quality level of the houses chosen by a given type of household is likely to be correlated with the observed attributes of these houses. Among the set of houses chosen by high-income households, for example, relatively small houses or houses with poor neighborhood amenities are likely to have high levels of unobserved house or neighborhood quality, leading to a downward bias in the estimation of the preferences of high-income households for the observed amenities. In fact, when the unobservable quality of each house is allowed to vary by type, the estimates of δ^t_h returned in the first stage of the estimation procedure are not consistent.[21] Consequently, we consider an alternative specification that deal with the does yield consistent estimates.

Type-Specific Neighborhood Unobservable Characteristics

In particular, we allow for the unobserved quality of each neighborhood to vary by type: $\eta^t_{h \in N}$, but force the valuation of the unobserved quality of each house within the neighborhood to be identical across households: ξ_h

$$(27) \qquad \underset{(h)}{Max} \quad V^{i,t}_h = \alpha^{i,t}_X X_h - \alpha^{i,t}_D D^i_h - \alpha^{i,t}_p p_h + \eta^t_{h \in N} + \xi_h + \varepsilon^{i,t}_h$$

[need to complete – main idea: including neighborhood sociodemographic characteristics interacted with household characteristics (e.g., percent highly educated interacted with household's education level) directly in analysis absorbs out $\eta^t_{h \in N}$.]

4 DATA

Our analysis 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 unlike the public-use data, also include information on the location of individual residences and workplaces at a very disaggregate level. In particular, while the public-use 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

[21] For a more complete discussion of the asymptotic properties of these models see Bayer, McMillan, and Rueben (2002).

individual inhabits as well as the characteristics of each neighborhood far more accurately than has been previously possible with such a large-scale data set.

Our study area consists of 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. First, it is reasonably self-contained. Examination of Bay Area commuting patterns in 1990 reveals that a very small proportion of commutes originating within these six counties ended up at work locations outside the area; and similarly, a relatively small number of commutes to jobs within the six counties originated outside the area. And second, 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 our data.[22] Our final sample consists of about 650,000 people in just under 244,000 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 (also provided at the Census block level).[23]

The Census data provide a variety of housing characteristics: whether the unit is owned or rented, the corresponding rent or owner-reported value, property tax payment, number of rooms, number of bedrooms, type of structure, and the age of the building. In constructing neighborhood characteristics, we begin by characterizing the stock of housing in the neighborhood surrounding each house. Using the Census data, we also construct neighborhood racial, education and income distributions based on the households within the same block group, a Census region containing approximately 500 housing units.[24] We merge additional data describing local conditions with each house record, constructing variables related to air quality, climate, 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

[22] Our sample consists of all households who filled out the long-form of the Census in 1990, approximately 1-in-7 households. In our sample, Census blocks contain an average of 6 households, while Census block groups – the next level of aggregation up - contain an average of 92 households.

[23] 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. The full list of the household characteristics used in the analysis, along with means and standard deviations, is given in the first panel of the Data Appendix.

[24] In principle, as we know the location of each house very precisely, neighborhoods could be defined to include all houses within a given radius of the house. In practice, the use of such measures yielded very similar results to those based on conventional Census boundaries, (e.g., Census blocks, block groups, or tracts), and consequently, we use these traditional Census boundaries when constructing neighborhood sociodemographic measures to facilitate comparison with past research. We discuss this issue further in Section 6 when describing the patterns of segregation in the Bay Area.

house is provided in a Data Construction Appendix.[25] The full list of house and neighborhood variables, along with means and standard deviations is given in the second panel of the Data Appendix.

Refining the House Price Variables Provided in Census

For a variety of reasons, the house price variables reported in the Census are ill suited for our analysis. House values are self-reported and top-coded, and rents may reflect substantial tenure discounts. Moreover, because we have implicitly defined the model and developed its equilibrium properties in terms of a single price variable for both owner-occupied and rental properties, we must relate house values to rents in some way.[26] Consequently, we make four adjustments to the housing price variables reported in the Census aiming to get a single measure for each unit that reflects what its monthly rent would be at current market prices. We describe the reasoning behind each adjustment here, leaving a detailed description of the methodology for the Data Appendix.

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 also 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

[25] In generating the climate and air quality data at the Census block level, for example, we make use of locally weighted regression techniques to assign data on climate stations and air quality monitoring stations to a lower level of aggregation (in this case, a Census block), as there are far fewer climate stations than Census blocks.

[26] This requirement may seem more restrictive than it actually is. Note that we treat ownership status as a fixed feature of a housing unit in the analysis. In this way, 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, will imply that wealthier households are more likely to own their housing unit, as we find below) and other house characteristics. Moreover, the model could incorporate heterogeneous elasticities of demand for features of a house or neighborhood depending on whether the unit is owned or rented. In this way, the use of a single house price variable does not impose any serious restrictions on the model.

for those individuals who appear to be reporting values much closer to the original transaction price rather than current market value.

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.

The third adjustment that we make concerns 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. 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.

Finally, we construct a single price vector for all houses, whether rented or owned. In order to make owner- and renter-occupied housing prices as comparable as possible, we seek to determine the implied current annual rent for the owner-occupied housing units in our sample. 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 these ratios to convert house values to a measure of current monthly rent. Again, the procedure is described in detail in the Data Construction Appendix.

School Characteristics

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 four 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.

The first three procedures do not make full use of the information available to us in the restricted version of the Census. Our preferred approach uses information about individual children living in each Census block - their age and whether they are enrolled in public school. (We are currently extending the procedure to use the race of each child as well.) We know the fourth grade enrollment for every school in every school district in the Bay Area. Adjusting for the fact that we have a sample of households in the long form of the Census, we know that the 'true' assignment of houses to schools must give rise to the overall fourth grade enrollments observed in the data. (It must also give rise to the observed racial composition of the school, a constraint we are not currently making sure binds.)

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, then we need to reassign Census blocks out of its catchment area, while if a school has excess supply, we need to 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 in practice, and produces plausible adjustments to the initial, closest-school assignment.

Boundary Fixed Effects

Three issues arise in incorporating boundary fixed effects into our analysis. The first issue concerns the choice of jurisdiction for which the boundaries are defined. While Black uses school attendance zones within a school district, in the analysis presented in this paper, we use boundaries between school districts in the Bay Area. 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 1 percent.[27] A second issue concerns the width of the boundaries. While a narrow band makes the assumption that unobserved neighborhood quality is the same on opposite sides of the boundary more accurate, a wider band allows the use of more data. To address this issue we consider a variety of alternative boundary widths in the analysis. A third issue concerns the length of the boundaries themselves. If a single fixed-effect were to be used for the Oakland-Berkeley boundary, for example, this boundary would be nearly 8 miles long. To address this issue, we again consider alternative boundary definitions that divide longer boundaries into much shorter segments.

5 RESULTS

[Tables to be handed out at talk – results not yet disclosed by Census for inclusion in paper]

6 CONCLUSION

REFERENCES

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

Bayer, Patrick, Robert McMillan, and Kim Rueben (2002) "The Causes and Consequences of Residential Segregation: An Equilibrium Model of Neighborhood Sorting," mimeo, Yale University, available at www.econ.yale.edu/~pjb37.

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.

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

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

Ekeland, Ivar, James Heckman, and Lars Nesheim, (2002), "Identification and Estimation of Hedonic Models," unpublished manuscript, 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.

[27] In the analysis presented below, we check the robustness of the use of school district boundaries by incorporating data that characterizes districts with special exemptions from this rule.

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.

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