DATA APPENDIX

1. Census Variables

House Prices. This section explains the construction of the house price variable used in our analysis, based on the self-report from the restricted-access version of the Census, combined with other Census and external data.

While the houses sampled in the Census have the advantage of being representative and the sample sizes are huge, the house values reported in the Census are subject to three potential problems: they are self-reported and may be subject to misreporting, they are tabulated in intervals, and they are top-coded. In light of these potential problems, we have generated a predicted house price measure using interval regression to deal with the categorical nature of the reported house value variable as well as the top-coding, and to refine the information contained within the self-report. Before describing the construction of the house price, we discuss the three potential problems briefly.

1. *Misreporting*

Because house values are self-reported in the Census, 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, asking 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 that exceeds US$ 7,000 at the time the current owner bought the property or in 1978 (whichever period is the most recent). 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 self-report.

2. *Tabulation in Intervals*

The coding of the house price variable in the Census involves restricting the variable to fall within one of 26 bands. For our purposes, a continuous point estimate is preferable. Because the property tax payment variable is continuous, it provides useful information in distinguishing the values of houses within intervals, in conjunction with a host of other housing and neighborhood characteristics available in the Census.

3. *Top-Coding*

House values reported in the Census are top-coded at $500,000, a restriction that is binding for many houses in California, even in 1990. 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.

House Price Measure

Using the self-reported values, we estimate interval regressions, which generalize the Tobit, separately for each of the 45 PUMAs in the Bay Area, restricting the house price point estimate to lie in the self-reported interval. In each case, we control for a number of housing characteristics, including the number of rooms, number of bedrooms, type of structure (single-family detached etc.), and age of the housing structure, as well as a series of neighborhood controls. We also include interactions of the property tax with tenure variables (in order to capture the effects of Proposition 13 on house prices), and interactions of the property tax, tenure variables and a dummy for the household head being 55 years of age or more (capturing the effects of Propositions 60 and 90 in California). We then calculate the predicted house values using the estimates from the interval regressions, conditional on being in the same interval as the self-reported value.

Rental Value

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 formal rent control is not in operation. Thus while this will not lead to errors in responding to the Census rental value 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 PUMA 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:

$$\log(R_j) = \beta_1 y_j + \beta_2 X_j + \upsilon_j \qquad (4)$$

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 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

In order to make owner- and renter-occupied housing prices comparable in our analysis, we need to calculate a current rental value for housing for both owned and rented units. Because house prices reflect 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, all included in X_j:

$$\log(\Pi_j) = \gamma_1 o_j + \gamma_2 X_j + \eta_j \qquad (5)$$

We estimate a series of hedonic price regressions of this form for each PUMA in 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 data we have added to the Census dataset, linked to Census blocks in our restricted-access data. These additional datasets include:

[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 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 procedures that take 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; the PAIF – a teacher-based form that provides detailed information about teacher experience, education and certification, and information on the classes each teacher teaches; and a language census that provides information on the languages spoken by limited-English-proficient students.

Procedures for Assigning School Data.

While we have an exact assignment of Census blocks to school attendance zones for around a third of the schools in the Bay Area, we employ an alternative approach to link each house to a school for our full sample. A simple procedure would assign each house to the closest school within the appropriate school district. Our preferred approach, which we use to generate the house-school match for our full dataset, 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 by matching 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 that school's *synthetic* attendance zone (recalling that we do not know the actual attendance zones for two-thirds of the schools in the Bay Area); in contrast, if a school has excess supply, we expand the school's attendance zones to include more blocks.

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 (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 Census block 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. For each Census block, we have calculated the percentage of land in ¼, ½,1, 2, 3, 4 and 5-mile radii 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, respectively.

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 of the national average etc. These ratings are provided by CAP index and were downloaded from APBNews.com.

Geography and Topography. The Teale Data Center provided information on the elevation, and latitude and longitude of each Census block.

TECHNICAL APPENDIX

Asymptotic Properties of the Estimator. Our sorting model fits within a class of models for which the asymptotic distribution theory has been developed. In this Technical Appendix, 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 (the number of housing types), N (the number of individuals in the sample), or C (the number of non-chosen alternatives drawn for each individual).[2]

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.

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 $H \rightarrow \infty$.[3] 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 (2004) 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 $H \rightarrow \infty$ 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 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

[2] As described in McFadden (1978), an attractive aspect of the IIA property for each individual is that we can estimate the multinomial logit model using only a sample, C, of the alternatives not selected by the individual. This permits estimation despite having many alternatives – i.e., many distinct house types.

[3] This condition requires certain regularity conditions. See Berry, Linton, and Pakes (2004) for details.

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 the 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 this requirement is met seems reasonable.