NBER WORKING PAPER SERIES

A UNIFIED FRAMEWORK FOR MEASURING PREFERENCES FOR SCHOOLS
AND NEIGHBORHOODS

Patrick Bayer
Fernando Ferreira
Robert McMillan

A Unified Framework for Measuring Preferences for Schools and Neighborhoods
Patrick Bayer, Fernando Ferreira, and Robert McMillan
NBER Working Paper No. 13236
July 2007
JEL No. H0,H4,H72,R0,R21,R31

ABSTRACT

This paper develops a comprehensive framework for estimating household preferences for school
and neighborhood attributes in the presence of sorting. It embeds a boundary discontinuity design
in a heterogeneous model of residential choice to address the endogeneity of school and neighborhood
attributes. The model is estimated using restricted-access Census data from a large metropolitan area,
yielding a number of new results. First, households are willing to pay less than one percent more in
house prices -- substantially lower than previous estimates -- when the average performance of the
local school increases by five percent. Second, much of the apparent willingness to pay for more educated
and wealthier neighbors is explained by the correlation of these sociodemographic measures with unobserved
neighborhood quality. Third, neighborhood race is not capitalized directly into housing prices; instead,
the negative correlation of neighborhood race and housing prices is due entirely to the fact that blacks
live in unobservably lower quality neighborhoods. Finally, there is considerable heterogeneity in preferences
for schools and neighbors: in particular, we find that households prefer to self-segregate on the basis
of both race and education.

Patrick Bayer
Department of Economics
Duke University
213 Social Sciences
Durham, NC 27708
and NBER
patrick.bayer@duke.edu

Fernando Ferreira
Wharton Real Estate Department
University of Pennsylvania
1461 Steinberg Hall-Dietrich Hall
3620 Locust Walk
Philadelphia, PA 19104
fferreir@wharton.upenn.edu

Robert McMillan
University of Toronto
Department of Economics
100 St. George Street
Toronto, ON M5S 3C3
CANADA
and NBER
mcmillan@chass.utoronto.ca

1   INTRODUCTION

Economists have long been interested in estimating household preferences for school and neighborhood attributes, given their relevance to many central issues in applied economics. Most fundamentally, preferences for schools and neighbors shape the way that households sort in the housing market,[1] influencing the level of residential segregation and the matching of households to schools. As such, reliable estimates of household preferences for schools and neighbors are essential in order to understand how schools, neighborhoods and houses are allocated in practice.

This paper develops a comprehensive framework for recovering household preferences for a broad set of school and neighborhood attributes in the presence of sorting. At its heart is a discrete choice model of the household residential location decision that allows household tastes to vary flexibly over housing and neighborhood characteristics. The model permits household choices to be influenced by unobservable choice attributes, and it nests two prominent frameworks for measuring household valuations for house and neighborhood characteristics – hedonic price regressions and traditional discrete choice models[2] – as special cases.

The paper's first main contribution is to provide a novel strategy for addressing the endogeneity of school and neighborhood attributes in the context of this heterogeneous sorting model. Of necessity, sorting correlates household and neighborhood attributes and in the process, induces correlations among a host of neighborhood attributes, including those that are unobserved. To account for the resulting significant endogeneity problems, we embed the *boundary discontinuity design* (BDD) developed by Black (1999) in our sorting model. Black's original application included school attendance zone boundary fixed effects in hedonic price regressions to control for the correlation of school quality and unobserved neighborhood quality.[3] In this paper, we show how the BDD can be extended in two key dimensions: first, to deal with the systematic correlation of neighborhood sociodemographic characteristics and unobserved neighborhood quality,[4] and second, to help identify the full distribution of household preferences for schools and neighbors.

---

[1] Intuition for the way sorting affects the housing market equilibrium derives from a long line of theoretical work in local public finance, following from Tiebout's seminal 1956 paper. Important contributions include research by Epple and Zelenitz (1981), Epple, Filimon, and Romer (1984, 1993), Benabou (1993, 1996), Fernandez and Rogerson (1996), and Nechyba (1997).

[2] Following Berry, Levinsohn, and Pakes (1995), we add a term that captures the unobserved quality of each residential choice, extending the traditional discrete choice model introduced by McFadden (1978).

[3] Intuitively, differences in house prices on opposite sides of school attendance zone boundaries reflect the discontinuity in the right to attend a given school, and therefore provide an estimate of the value that households place on the difference in school quality across the boundary.

[4] Because of the inherent difficulty of isolating variation in neighborhood sociodemographics uncorrelated with unobserved aspects of neighborhood and housing quality, many researchers – see Cutler, Glaeser, and Vigdor (1999) and Bajari and Kahn (2005), for example – have simply elected to recognize the endogeneity

Based on our sorting model and its extended boundary identification approach, the paper's second main contribution is to provide new estimates of household preferences for schools and neighbors. To that end, we make use of a unique dataset, built upon a restricted-access version of the U.S. Census, that links detailed characteristics for nearly a quarter of a million households and their houses in the San Francisco Bay Area with their precise residential location (down to the Census block). This precise matching of households to their houses and neighborhoods allows us not only to estimate the heterogeneous sorting model but also to characterize detailed variation in sociodemographic characteristics on a block-by-block basis.

To motivate our general framework, we begin with a descriptive analysis of sorting at school attendance zone boundaries using these rich Census data. Given a discontinuity in local school quality at a school boundary, one might expect that residential sorting would lead to discontinuities in the characteristics of households residing on opposite sides of the same boundary; even if a school boundary was initially drawn such that the houses immediately on either side were identical, households with higher incomes and education levels might be expected to sort onto the side with the better school. This consequence of sorting is clearly apparent in our empirical analysis: nonparametric plots in the region of school attendance zone boundaries show sharp changes in household income, education, and race, with higher-income, better-educated households sorting onto the side of the boundary with higher school quality. At the same time, housing characteristics are more or less continuous.

In an idealized setting – one in which researchers were able to compare a vast number of houses facing each other directly but on opposite sides of the same boundary – these differences in sociodemographics would be of little import: the neighborhoods experienced by households on each side of the boundary would, to all intents and purposes, be the same.[5] In practice, researchers are forced to compare houses in bands – often over 0.3 miles wide – on either side of school boundaries in order to have sufficient sample sizes for inference.[6] Given the clear differences in sociodemographics that arise through sorting, it then becomes potentially important to control for differences in neighbor characteristics, as the house price differences found in the recent boundary fixed effects literature may reflect not only the discontinuity in school quality,

---

of neighborhood sociodemographics as a limitation of their analysis, having no reasonable way to address it. In other cases, researchers have isolated variation in neighborhood sociodemographics within Census tracts or other broader regions, though the underlying factors causing variation in sociodemographics are unobserved, and thus the fundamental endogeneity problem described here remains.

[5] We note, however, that *school* peers would differ discontinuously right at the boundary. In the analysis below, we are able to account for such differences in school peers, with little effect on our main findings.

[6] Black (1999) compares results from three subsamples: 0.35 miles, 0.2 miles and 0.15 miles to the nearest boundary, while Kane *et al.* (2003) focus on houses within 2000 feet, 1000 feet and 500 feet of the closest boundary, corresponding to 0.4, 0.2 and 0.1 mile bands respectively.

but also the value that households place on the corresponding differences in the characteristics of their immediate neighbors.[7]   As in Black (1999), our results indicate that the inclusion of boundary fixed effects substantially reduces the coefficient on school quality in hedonic price regressions. But the subsequent inclusion of precise neighborhood sociodemographic controls reduces this estimate further, by approximately 50 percent, even when constraining the sample to narrower bands of 0.10 mile.

Next, we show that the boundary approach can be extended to learn about household valuations of neighborhood sociodemographics. Our key insight is that household sorting across boundaries generates variation in neighborhood sociodemographics that is primarily related to an *observable* aspect of neighborhood quality – in this case, schools. Thus, to the extent that one can control for differences in school quality on opposite sides of the boundary, a boundary discontinuity design provides a plausible way to estimate the value that households place on the characteristics of their immediate neighbors.[8] In a hedonic price regression setting, we show that the inclusion of boundary fixed effects reduces the magnitudes of the coefficients on the income and education of one's neighbors by 25 and 60 percent, respectively. This is consistent with the intuitive notion that higher-income and better-educated households select into neighborhoods with better amenities. Even more strikingly, the magnitude of the coefficient on the fraction of black neighbors declines to zero. This implies that the negative correlation of housing prices and fraction of black neighbors observed in our dataset, and reported systematically in the previous literature, is driven by the correlation of race and the unobserved neighborhood quality captured by the boundary fixed effect.

In general, it is difficult to determine how the estimates of hedonic price regressions relate to fundamental preferences in the population. Using our sorting model, we show that if households are homogeneous, estimation reduces to a hedonic price regression, consistent with the notion that market prices must reflect mean preferences when all households are identical. When households are heterogeneous, estimates of a hedonic price regression need not return mean preferences. In this case, our sorting model provides an intuitive adjustment to the hedonic

---

[7] The regression discontinuity design (RDD) literature notes identification problems arising from sorting, since the quasi-random assignment of treatment and control groups in an RDD becomes invalid once individuals self-select into the treatment. For a recent exposition of this problem, see Lee (2007), although the older RDD literature also makes this point clear – see Cook and Campbell (1979), for example.

[8] Our identifying assumption is that the included controls for neighborhood sociodemographics – percent highly educated, average income, percent black, Hispanic and Asian, respectively – capture everything relevant about the characteristics of one's neighbors. This assumption would be necessary in any circumstance where one wanted to estimate value of neighborhood amenities. Unlike controlling for fixed effects at a broader geographic level, where the variation in neighborhood sociodemographics is still systematically related to unobserved aspects of housing and neighborhood quality, our approach gives us a handle on the fundamental source of sorting at these boundaries.

price regression accounting for differences in valuation between the mean and marginal household.

Estimates of the general sorting model using our rich dataset indicate that the hedonic price regression coefficients are generally very close to mean preferences for housing and neighborhood attributes that vary more or less continuously throughout the metropolitan area, including school quality and neighborhood income and education. In contrast, estimated mean preferences for neighborhood race differ significantly from the coefficients of the hedonic. We find that estimated mean preferences for black neighbors are significantly negative, differing markedly from the hedonic estimates, reflecting the fact that blacks make up less than 10 percent of the population so that mean and marginal households are far apart. The estimates of our sorting model also indicate that there is considerable heterogeneity in preferences for schools and neighbors. Perhaps most interestingly, our analysis implies that, conditional on neighborhood income, households prefer to self-segregate on the basis of both race *and* education.

It is important to underscore some limitations of our approach. First, the sorting model only deals with preference heterogeneity that varies with observable household characteristics. Although our Census dataset allows the inclusion of a large number of observable features of each household and housing unit, future research could potentially adopt a random coefficients specification.[9] Second, the empirical strategy adopted in this paper takes into account a number of important endogeneity concerns, but it does not address the possibility that the higher-income households on the higher test score side of a school boundary might be more likely to make home improvements (install granite countertops, for example) unobserved by the researcher, in turn contributing to the higher average house prices on that side of the boundary. That said, we are unaware of any paper in the literature that has been able to deal with this issue.

The rest of the paper is organized as follows: In Section 2, we describe the data used in the analysis. Descriptive evidence on sorting at school attendance zone boundaries is presented in Section 3, and hedonic estimates, in Section 4. The sorting model is set out in Section 5, our estimates of the model are discussed in Section 6, and Section 7 concludes.

2 DATA

Census Dataset

The primary dataset used in our analysis is drawn from the restricted-access version of the 1990 Decennial Census. This dataset provides information for the full sample of households who filled out the long-form questionnaire – approximately 15 percent of the population. For

---

[9] This would come at a cost, though, as additional structure is needed to estimate unobserved heterogeneity.

each household, these data provide a wide range of economic and demographic variables, including the race/ethnicity, age, educational attainment, and income of each household member. In addition, the data also characterize each household's residence: 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.

For our purposes, the most important feature of this restricted-access Census dataset is that it characterizes the location of each individual's residence and workplace very precisely; these locations are specified at the level of the Census block (a region with approximately 100 individuals) rather than the publicly-available Census PUMA (a region with an average of 100,000 individuals). This precise geographic information allows us to examine the way that households and houses vary from block-to-block anywhere within our study area.

The study area for our analysis 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: a very small proportion of commutes originating within these six counties in 1990 ended up at work locations outside the area, and vice versa. Second, the area is sizeable along a number of dimensions: it includes over 1,100 Census tracts, 4,000 Census block groups, and almost 39,500 Census blocks, the smallest unit of aggregation in the data. Our full sample consists of around 650,000 people in 242,100 households.

For this sample, we construct a variety of housing and neighborhood variables based on the restricted-access Census data. We use information provided by the head-of-household to construct a predicted house price measure. Renters simply report a measure of the current monthly rent, while owners report an estimate of the current market value of the house,[10] and we place house values and rents on the same monthly basis to obtain a single house price variable.

We also construct a set of detailed neighborhood-level variables characterizing the racial, education and income composition of each Census block and Census block group. We merge additional local data with each house record, relating to crime rates, land use, topography, urban density, and local schools.[11] As our primary measure of school quality, we use the average fourth grade mathematics and reading test score for each school, averaged over two years, this averaging

---

[10] We refine the self-reported house value variable so as to reduce some of the measurement error in it. In particular, the house value variable recorded in the Census is a categorical variable, falling into one of 26 bins, including a bin for top-coded values ($500,000 or more in 1990). Using additional information on a continuous measure of property taxes and a rich set of house and neighborhood controls, along with the rules implicit in Proposition 13, we convert this categorical variable into a point estimate for each housing unit. (See the Data Appendix for a fuller discussion.)

[11] See the Data Appendix for more details.

helping to reduce any year-to-year noise in the school quality variable. While the average test score is an imperfect quality measure, it has the advantage of being easily observed by both parents and researchers; as a result, it has been used in most analyses that attempt to measure the demand for school quality.[12]

Summary statistics for the primary housing and neighborhood variables in our full sample are given in the first two columns of Table 1 (and repeated in Table 2).[13] In the 1990 Census, average house values in this sample are around $300,000 and rents, approximately $750 per week. The average test score, our measure of school quality, has a mean of 527 and a standard deviation of 74. Around 60 percent of homes are owned and the average number of rooms per housing unit is just over five. In terms of neighborhood sociodemographics, Census block groups in our full sample are on average 68 percent white and 8 percent black; 44 percent of the heads of household in each Census block group have a college degree or more, and average block group income is just under $55,000.

Transactions Dataset

As a complement to the restricted-access Census data, we have also assembled a dataset that characterizes the complete set of housing transactions in the San Francisco Bay Area between 1992 and 1996. These data are based on County public records, and contain detailed information about every housing unit sold during that period, including the exact transaction price and the exact street address.[14] We use the transactions data to investigate the robustness of our findings, given that Census housing prices are self-reported, though we note that this dataset is not representative of the full sample of households – the stock of homeowners and renters – living in a neighborhood, instead capturing the flow of new homeowners into a neighborhood.

While our transactions dataset does not directly include demographic information on home buyers, we were able to add some buyer characteristics by drawing on data collected in accordance with the Home Mortgage Disclosure Act (HMDA). Enacted by Congress in 1975 and implemented by the Federal Reserve Board's Regulation C, the HMDA data provide some description of the buyer/applicant (including household income), as well as the mortgage loan amount, the mortgage lender's name, year of the transaction, and the Census tract where the

---

[12] In specifications designed to study the robustness of our baseline results, we also include other schooling measures that characterize the school's teachers and peers.

[13] The full sample of 242,100 households is used in the first step of the estimation of the logit model, described in Section 5. Our boundary subsamples, summarized in columns (3) – (7) of Tables 1 and 2, are used to study sorting at school attendance zone boundaries – see Sections 3 and 4 – and in the second step of the estimation procedure.

[14] Black (1999) used a similar housing transactions dataset from the Boston area.

property is located.[15] We were able to merge the HMDA data with our housing transactions on the basis of Census tract, loan amount, date, and lender name. This procedure resulted in unique matches for approximately 60 percent of all home sales, and allowed us to generate neighborhood variables for 85 percent of the sample. The first column of Appendix Table 1 presents a description of this sample, which we use in the regressions below to help gauge the robustness of our main findings.

School Attendance Zone Boundaries

In order to implement the boundary approach, we gathered school attendance zone maps for as many elementary schools as possible in the Bay Area, for the period around the 1990 Census.[16] Our final attendance zone sample consists of 195 elementary schools – just under a third of the total number in the Bay Area. From this sample, we excluded portions of boundaries coinciding with school district boundaries, city boundaries, or large roads, since they could potentially confound our identification strategy.

For Census blocks falling within these attendance zones, we followed a simple procedure to assign each block to a boundary. For each block, we calculated the perpendicular distance from the block population centroid to the nearest school attendance zone boundary. We then located the closest 'twin' Census block on the other side of that boundary. If a given block had a lower score than its twin, it was designated as being on the 'low' side of the boundary; otherwise it was designated as being on the 'high' side of the boundary. We restrict attention to boundaries for which we have Census data on both high and low sides.

For our main boundary analysis, we focus on houses in all Census blocks that are within 0.20 miles of the closest school attendance zone boundary. The average distance to the boundary for this subsample is thus quite a lot smaller than 0.20 miles. For comparison, we also analyze a further subsample, consisting of houses assigned to Census blocks within 0.10 miles of the closest attendance zone boundary. Although the 0.10-mile subsample includes approximately half the number of observations, it provides a closer approximation to the ideal comparison of houses on the opposite sides of the same street, though in separate attendance zones.

---

[15] The Act requires lending institutions to report public loan data. Its purpose is to provide public loan data that can be used to determine whether financial institutions are serving the housing needs of their communities and whether public officials are distributing public-sector investments so as to attract private investment to areas where it is needed. The data are also intended to help identify any possible discriminatory lending patterns. (See http://www.ffiec.gov/hmda for more details).

[16] School attendance zone maps are not provided or catalogued by the State of California. Therefore, we contacted all local school districts and schools individually and requested detailed maps for each school attendance zone within a district during the period of analysis. Subsequently, these maps were digitized in order to be used in this research.

Column 3 of Table 1 shows averages for the 0.20 miles subsample, and Column 3 of Table 2 presents analogous numbers for the 0.10 miles subsample. When comparing these to the full Bay Area sample (column 1), it is clear that prices, test scores, ownership, house size, average income and percentage white are slightly lower in the boundary subsamples. This is due in large part to the absence of San Francisco from our boundary samples, given that it does not have well-defined attendance zones.

3   DISCONTINUITIES AT ATTENDANCE ZONE BOUNDARIES

In this section, we present descriptive evidence that sheds light on household sorting in the region of school attendance zone boundaries. We take advantage of the block-level information provided in the restricted version of the Census to measure the characteristics of housing units and households in a precise way on each side of a given boundary. Throughout this section, we focus on boundaries for which the test score gap comparing low and high sides is in excess of the median gap (38.4 points); if schools were identical on either side, there would be little reason to expect to see sorting.

We begin with a series of figures that summarize the movement of variables in the boundary region. The figures are constructed by first regressing the variable in question on boundary fixed effects and on distance-to-the-boundary dummy variables, then plotting the coefficients on these distance dummies. Thus a given point in each figure represents the conditional average (in 0.02 mile bands) of the variable in question at a given distance to the boundary, where negative distances indicate the 'low' test score side; all averages are normalized to zero at the closest point on the low side of the boundary.

By construction, and as shown in top left panel of Figure 1, there is a clear discontinuity in average test score at the boundary. For the Census sample considered, the magnitude of the discontinuity is around 75 points (which is approximately a standard deviation). The top right panel of Figure 1 shows a similar pattern for the test scores assigned to the housing transactions data. The bottom left panel of Figure 1 shows the difference in house prices on low versus high sides using the Census data, which corresponds to approximately $18,000 at the threshold. The more precise transaction price data in the bottom right panel shows a similar seam: a $20,000 difference right at the boundary.

As Black (1999) pointed out, if all housing and neighborhood amenities were continuous at the boundary, then these differences in price would correspond to the observed gap in school quality. Given the proximity of houses across the boundary, it is probably reasonable to expect a

somewhat similar housing stock at the threshold.[17] We test this assumption by comparing housing characteristics across the boundary. The panels of Figure 2 show that the housing variables drawn from the Census – average number of rooms, ownership, and year built – are continuous through the boundary. Similarly, Figure 3 shows that the housing variables in our transactions dataset are also reasonably continuous through the boundary, perhaps with the exception of square footage, though we note that transactions data are less-representative, consisting of a sample of recently moved-in homeowners.

In contrast, Figure 4 presents a different story with respect to the people inhabiting those houses. On average, the households on the high test score side of the boundary have more income and education, and are less likely to be black. This observed sorting at attendance zone boundaries provides initial evidence suggesting that household preferences for schools are heterogeneous.

The corresponding statistical tests for the presence of discontinuities at the boundaries are reported in the final column of Tables 1 and 2 for the 0.20- and 0.10-mile subsamples, respectively, using the same subsamples as the figures. For each subsample, the tests are based on regressions of the running variable on boundary fixed effects and a dummy that designates the high side of the boundary, clustering at the school attendance zone level. These tests underscore the main findings from the figures: that test scores and house prices are discontinuous at the boundary, that housing attributes are reasonably continuous in this sample, and that neighborhood sociodemographics present a significant seam at the attendance zone boundaries. Interestingly, we also find minimal evidence of a discontinuity in monthly rents, which suggests that average test scores and neighborhood characteristics are capitalized more fully into property values than rents. Appendix Table 1 reports similar tests using the housing transactions data. The results for this more select sample broadly corroborate the Census findings.

Collectively, the results presented in this section indicate that house prices respond positively to the variation in test scores across the boundaries. Moreover, the results also clearly indicate that households sort with respect to school attendance zone boundaries. These descriptive results have two immediate consequences for hedonic analyses. First, because sorting with respect to boundaries is pronounced, ignoring it is likely to lead to an overstatement of demand for schools versus the characteristics of immediate neighbors. This issue is likely to be especially relevant for analyses that include houses not in the immediate vicinity of a boundary. Second, the significant variation in neighborhood as well as school sociodemographic

---

[17] It is important to keep in mind that these school attendance zone boundaries are not school district boundaries, not city boundaries, and not aligned with major roads.

characteristics in the boundary region suggests that a boundary discontinuity design may prove useful in learning about willingness-to-pay for neighborhood sociodemographic characteristics. We now explore these consequences further.

4   HEDONIC PRICE REGRESSIONS

In this section, we use a regression framework to investigate relationships among key variables in the region of school attendance zone boundaries.   Doing so brings to light consequences for the boundary identification approach that have not been addressed in prior research.   In particular, we show that controlling for neighborhood sociodemographics has a quantitatively significant effect on the school quality coefficient in hedonic price regressions, even when accounting for neighborhood unobservables.   We also show that the negative correlation between house prices and neighborhood race widely reported in the literature is fully explained by the correlation between neighborhood race and unobserved neighborhood quality.

Our main estimating equation relates the price of house h to a vector of housing and neighborhood characteristics X_h and a set of boundary fixed effects, θ_{bh}, which equal one if house h is within a specified distance of boundary b and zero otherwise:

$$(1) \quad p_h = \beta X_h + \theta_{bh} + \xi_h$$

To maximize the sample size in our baseline analysis, we include both owner- and renter-occupied units in the same sample.   To put these units on a comparable basis, we convert house values to a measure of monthly user costs using a hedonic regression that returns the average ratio of house values to rents for housing units with comparable observable characteristics; we do so for each of 45 sub-regions of the Bay Area.[18]

A comparison between hedonic regressions for the full sample versus the houses in the boundary subsamples (within 0.20 and 0.10 miles respectively) is shown in Appendix Table 2. All estimated coefficients for test scores and neighborhood sociodemographics indicate that hedonic results *without* the inclusion of boundary fixed effects hardly change when constraining the sample to narrow bands around attendance zone boundaries.   From this point on, we restrict

---

[18] Separate estimation for each sub-region (a Census PUMA) allows the relationship between house values and current rents to vary with local expectations about the growth rate of future rents in the market.   The average estimate of the ratio of house values to rents is 264.1.   In subsequent analysis, we report estimates of specifications of equation (1) that limit the sample to only owner-occupied units and use the original house value variable as the dependent variable.   See additional details in the Data Appendix.

our attention to houses located with 0.20 and 0.10 miles of a boundary line, since the boundary fixed effects are only defined for houses located in those areas.

Table 3 reports estimates for the key parameters for a total of eight specifications of this hedonic price regression, using the monthly user cost of housing as the dependent variable. The reported specifications differ along three dimensions: (i) whether neighborhood sociodemographics are included in the specification, (ii) whether boundary fixed effects are included, and (iii) whether the sample consists of houses within 0.20 miles versus 0.10 miles of a boundary. All of the specifications include a full set of controls for housing and neighborhood characteristics, which are listed in the table notes.

Baseline Results for Average Test Score

In discussing the results in Table 3, we focus first on the specifications reported in the first two columns of the upper panel, labeled (1) and (2). These use the sample of houses within 0.20 miles of a school attendance zone boundary and exclude neighborhood sociodemographic measures. Comparing the estimated coefficients on average test score in these specifications, the results are qualitatively and quantitatively similar to those reported in Black (1999). In particular, the estimated effect of a one standard deviation increase in a school's average test score on the cost of housing declines by nearly 75 percent, from $124 to $33 per month, when boundary fixed effects are included in the analysis. This suggests (as in Black) that the majority of the observed correlation between test scores and housing prices is driven by the correlation of school quality with other aspects of housing or neighborhood quality.[19]

Continuing to focus on the first two columns of Table 3, we next compare the estimated coefficients on average test score in the upper versus the lower panel. This comparison highlights the additional impact of controlling for neighborhood sociodemographic characteristics, over and above the inclusion of boundary fixed effects. Estimates from the column labeled (4) show that the addition of detailed sociodemographic measures reduces the coefficient on average test score to $17 per month.[20] This reduction is due entirely to the sorting of households across school attendance zone boundaries already shown in the descriptive tables and figures above. The magnitude of this reduction – 50 percent – highlights the fact that the inclusion of boundary fixed

---

[19] Black (1999) finds that a 5 percent increase in test scores change house prices by 4.9 percent for the full sample, and only by 2.1 percent when controlling for boundary fixed effects.

[20] The low estimated value may partly reflect the informational problem households face in attempting to distinguish the quality of a school.

11

effects in a hedonic price regression is not fully effective in controlling for all aspects of neighborhood quality.[21]

Our preferred estimate of $17 per month for a one standard deviation increase in the average test score is roughly 1.8 percent of the average monthly user cost of housing and corresponds to approximately $4,500 in house value in 1990. The key assumptions underlying the interpretation of this estimate as the market value of school quality are: (i) that unobserved housing characteristics do not vary across the boundary, and (ii) that the measures for neighborhood race/ethnicity, education, and income included in regression control fully for sorting across boundaries. There is also a possibility that the average test score captures something else about the school (e.g. peers or teachers) that households actually value. We explore this issue further in the robustness section below.

Baseline Results for Neighborhood Sociodemographic Characteristics

Comparing the coefficients on neighborhood sociodemographic characteristics in the specifications shown in columns (3) and (4) of Table 3 provide an estimate of the bias associated with the sorting of higher-income and better-educated households into neighborhoods with different levels of unobserved neighborhood quality. In particular, the inclusion of boundary fixed effects leads to a 25 percent decline in the coefficient on the average income of one's neighbors, from $60 to $45 per month (for a $10,000 increase), and a 60 percent decline in the coefficient on the fraction of neighbors that are college-educated, from $220 to $90 per month. These results suggest that analyses which fail to control for the correlation of neighborhood sociodemographics with unobserved neighborhood quality are likely to significantly overstate the extent to which neighborhood socioeconomic characteristics are capitalized into property values.

The effects for neighborhood race are perhaps even more interesting. With the inclusion of boundary fixed effects, the coefficient on the percent of one's neighbors who are black changes from -$100 to $2. This implies that the racial composition of a neighborhood is not capitalized directly into housing prices; instead, the large negative correlation of housing prices and the fraction of black households in a neighborhood reflects in its entirety the correlation of unobserved aspects of neighborhood quality with neighborhood race. This empirical finding is, to the best of our knowledge, new to the literature. While many prior studies have documented the correlation of race and housing prices, ours is the first to use a boundary discontinuity design to address the correlation of neighborhood race and unobserved neighborhood quality.

---

[21] It is also worth noting that, for this sample, controlling carefully for neighborhood sociodemographic characteristics has *as large* an impact on the coefficient on school quality as the inclusion of boundary fixed effects, in both cases reducing the point estimate to $33-$35 per month.

The key advantage of using a boundary discontinuity design to estimate the market value of neighborhood sociodemographics is that it isolates variation in these characteristics that is primarily related to an *observable* aspect of neighborhood quality. Thus, under the standard assumptions of the BDD, as well as the assumption that the school characteristics included in the regression – average test score and, in additional specifications below, measures related to peers and teachers – control fully for differences in school quality, so the variation in neighborhood sociodemographics at boundaries is uncorrelated with the unobservable. In contrast, a research design that isolates variation in neighborhood composition across blocks or block groups within a broader geographic level (e.g., through the inclusion of Census tract fixed effects) continues to rely on variation in neighborhood composition systematically related to any differences in the unobserved aspects of housing and neighborhood quality across blocks or block groups.

As we discuss in greater detail in Section 5 below, the statistically and economically insignificant coefficients on neighborhood race in specification (4) by no means imply that households do not have strong racial preferences – on the contrary, the heterogeneous preferences we estimate in the sorting model indicate that households have strong self-segregating preferences. Rather, the fact that race is not capitalized into housing values suggests that households are able to sort themselves across neighborhoods on the basis of race without the need for price differences to clear the market. We return to this issue once we have reported estimates from the heterogeneous sorting model below.

Robustness Checks

To examine the robustness of our main findings and also to help distinguish among alternative explanations for the patterns we have described, we now consider how the results presented in the columns (1) through (4) of Table 3 compare with analogous specifications.

A. Distance to the Boundary

As described in Black (1999), the idealized use of a boundary discontinuity design would compare the prices of houses on opposite sides of a neighborhood street that served as a boundary between school attendance zones. Such a comparison would hold everything about the neighborhood as close to constant as possible, and any discontinuity in house prices would be almost completely attributable to differences in the valuation of the assigned schools. In reality, in order to generate large enough samples, researchers employing a BDD have typically used a sample of houses within a threshold distance of a boundary in the range of 0.15-0.35 miles.

Due of the size of our dataset, we are able to consider a threshold distance of 0.10 miles, rather than 0.20 miles, to the closest school attendance zone boundary. These results are reported

in columns (5) through (8) of Table 3. Comparing these coefficient estimates to those for the 0.20-mile sample makes clear that the qualitative nature of the findings described remains unchanged using the smaller sample. As this pattern holds more broadly, we focus on results using the 0.20-mile sample in the remaining tables both because these tend to be more precise and to avoid redundancy.[22]

B. School Characteristics versus Immediate Neighbors

One explanation for our baseline results is that individuals have preferences over their immediate neighbors (e.g. the individuals who reside on just the same block instead of within a broader surrounding circle) and that even at a threshold distance of 0.10 miles, these vary significantly enough to matter. An alternative explanation is that households value school sociodemographic characteristics over and above school quality as reflected in the average test score. In this case, the neighborhood sociodemographic measures included in our baseline specification might proxy for school-level differences.

Specifications (A)-(C) in Tables 4 and 5 are designed to examine the role of these alternative explanations for our baseline results for average test score and neighborhood sociodemographics respectively. Specification (A) adds a series of controls that characterize the race, language ability, and income of the children in the elementary school, as well as the average education of the teachers.[23] As can be seen in the tables, the inclusion of these well-measured school controls does little to change the pattern of results for either the coefficient on average test score (Table 4) or the set of coefficients on the included neighborhood sociodemographic measures (Table 5). Thus, households do not seem to place significant value on the variation in school sociodemographics that is not explicitly correlated with either the average test score or local neighborhood sociodemographics.[24]

That the inclusion of *school* sociodemographic measures does little to affect the analysis suggests that the valuation of one's immediate neighbors may be an important factor explaining the significant impact of controlling for neighborhood sociodemographics. To explore this

---

[22] A full set of results for the 0.10-mile sample is available from the authors upon request.

[23] In particular, controls are included for the fraction of Asian, Black, and Hispanic children in the elementary school, the fraction of limited English proficiency, and the fraction receiving free lunch. An additional variable measures the fraction of teachers whose educational attainment does not exceed a bachelor's degree.

[24] One explanation for this result is that households may sort on the basis of published test scores and neighborhood sociodemographics. This would be natural if households found it difficult to separate out the portion of the test score attributable to school sociodemographic composition from the underlying effectiveness of the school. Rothstein (2006) addresses this issue. Instead of modeling residential location and schooling decisions, he uses variation across school districts applied to a set of 1994 SAT-takers in a bid to disentangle parental choice based on school effectiveness and peer groups respectively. His findings suggest that parents have difficulty distinguishing these components.

possibility more fully, specification (B) reports results for a regression that includes controls for neighborhood sociodemographics measured at the Census block level, along with our baseline neighborhood demographic measures (measured at the block group level). The coefficients on average test score reported in Table 4 change very little relative to our baseline results with the inclusion of block-level controls. This suggests that our baseline measure, which uses the average composition of the portion of the block group on the same side of the boundary, does a reasonably good job of capturing the variation in immediate neighbors across boundaries.

The corresponding coefficients on both sets of sociodemographic measures reported on Table 5 imply that households do indeed place significant value on the education and income levels of their immediate neighbors – those on the same block.[25] In particular, the results reported for specification (B) indicate that, *conditional on block group composition*, a 10 percent increase in the fraction of college-educated neighbors on the same block raises house prices by an additional $6 per month, and a $10,000 increase in the average income of households in the same block raises house prices by an additional $25 per month. Neighborhood race continues to have an insignificant effect on housing prices in these hedonic price regressions.[26]

Specification (C) addresses a further robustness issue related to the construction of the block group neighborhood characteristics. In this case, rather than limiting the measure to the portion of the block group on the same side of the boundary, we include standard block group averages that may span the boundary for block groups very close to it. While the results are qualitatively similar to the pattern of results already shown, the impact of controlling for neighborhood sociodemographics on the average test score coefficient is dampened by a small amount. This result is not entirely surprising given that this method of assigning neighborhood sociodemographic characteristics systematically averages the block group level measures across boundaries.

C. Top-coding of Census Prices

Specification (D) in Tables 4 and 5 considers a sample in which top-coded houses (those with values equal to or greater than $500,000 in 1990) are dropped from the sample. While the magnitudes of the coefficients on average test score are smaller when boundary fixed effects are not included in the analysis, the results are nearly identical when fixed effects are included.

---

[25] In reading these results, it is important to keep in mind that the coefficients on the block-level measures capture the *additional* impact of variation at the block over and above the contribution this variation makes to the block group-level measures.

[26] That a sizeable portion of the effect of controlling for neighborhood sociodemographic measures is attributable to the capitalization of the characteristics of immediate neighbors into housing prices is also broadly consistent with the non-parametric plots of house values and neighborhood sociodemographics in the region of school attendance zone boundaries shown in Figures 1 and 4.

D. Only Owner-Occupied Units

Specifications (E)-(F) in Tables 4 and 5 restrict attention to owner-occupied units based on samples drawn from the Census and our transactions dataset, respectively. Specification (E) reports coefficients for a specification in which the dependent variable is the house value reported directly in the Census rather than the monthly user cost of housing that we use in our main analysis. While the qualitative pattern of results for the owner-occupied units mirrors that of the full sample, the results for average test scores are substantially greater in magnitude. With the inclusion of boundary fixed effects and neighborhood sociodemographic characteristics, a one standard deviation in average test scores is associated with a $9,400 increase in property values (the mean property value in the 0.20-mile boundary sample is $250,000). This is equivalent to approximately $35 in monthly user costs – which is roughly twice the baseline estimate shown in the first row.[27] The corresponding coefficients on the neighborhood sociodemographics reported in Table 5 continue to suggest that neighborhood income is the characteristic most directly capitalized into property values – the coefficients on both neighborhood race and education are statistically insignificant when boundary fixed effects are included in the analysis.

Specification (F) is based on actual transactions observed in our transactions dataset. This specification allows us to gauge the robustness of our findings with respect to the self-reported house values in the Census. Before discussing particular parameter estimates, it is worth re-emphasizing a number of key differences between this sample and that based on the restricted Census data. First, the housing prices are based on actual transaction prices rather than self-reported values from Census respondents. Second, the sample includes not only owner-occupied houses, but also restricts attention to those who have had a recent transaction within a reasonably small window near the Census year. Finally, the only neighborhood sociodemographic measure included, average income, is based on a sample of recent transactions and thus reflects the characteristics of the flow of households currently moving into a neighborhood rather than the stock of households already residing in the area.

Despite these differences, the results reveal a strikingly similar pattern to those reported in specification (E). The coefficient on the average test score declines by nearly 65 percent with the inclusion of boundary fixed effects (from $34,000 to $12,000 in house value) and then declines another 33 percent (to $9,000) with the inclusion of neighborhood sociodemographics. The coefficient on average neighborhood income, which proxies for all neighborhood sociodemographics in this specification, declines more than 55 percent (or from $15,800 to

---

[27] Consistent with this conclusion, price regressions estimated on the sample of renters reveal a slightly positive and insignificant coefficient on average test score when boundary fixed effects are included in the analysis.

$6,800 for a $10,000 increase in average income). Thus, our baseline analysis appears to be robust to the use of house values based on the self-reports from the Census.

Summary

Overall, the qualitative pattern of results is remarkably robust across the full set of specifications reported in Tables 4 and 5. Three main conclusions emerge. First, sorting at school attendance zone boundaries is an important phenomenon, already clear from the earlier graphical analysis. Second, even when boundary fixed effects are included in the analysis, failing to control for such sorting leads to a significant overstatement of the capitalization of average test scores into house prices. Third, controlling for differences in unobserved neighborhood quality using a boundary discontinuity design leads to a substantial reduction in the estimated effect of neighborhood socioeconomic and (especially) racial characteristics on housing prices.

5   SORTING MODEL

The clear evidence of sorting across school attendance zone boundaries naturally suggests that households vary in their willingness to pay for at least some features of schools and neighborhoods. This raises the obvious issue of how the coefficients in the hedonic price regressions reported in Section 4 relate to underlying household preferences. In this section, we develop a heterogeneous model of residential sorting, using boundary fixed effects to help identify the entire distribution of preferences for schools and neighborhoods. The model clarifies the relationship between the distribution of preferences and the hedonic price of school and neighborhood characteristics – in particular, when the coefficients in a hedonic price regression are likely to provide a reasonable approximation to the mean marginal willingness-to-pay of the population and when they are not.

Model

We model the residential location decision of each household as a discrete choice of a single residence.[28] The utility function specification is based on the random utility model developed in McFadden (1973, 1978) and the specification of Berry *et al.* (1995), which includes

---

[28] Following McFadden (1978), a long line of papers use discrete choice models to estimate residential choice. Many of these papers, including Quigley (1985), Nechyba and Strauss (1998), and Barrow (2002), focus specifically on estimating preferences for school quality. A related line of research using hedonic demand models, including Rosen (1974), Epple (1987), Bajari and Benkhard (2002), Ekeland, Heckman, and Nesheim (2004) and Heckman, Matzkin, and Nesheim (2003), provides an alternative approach to estimating demand for non-marketed goods and attributes. The fundamental difference between hedonic demand and discrete choice models is that the former assume that households are able to select the level of consumption of each attribute to satisfy the relevant first-order condition, while the latter explicitly account for the fact that households are constrained to choose among the finite set of choices in the data.

choice-specific unobservable characteristics. 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 the surrounding neighborhood (e.g., school, crime, population density, and topography). Let p_h denote the price of housing choice h and let d_h^i denote the distance from residence h to the primary work location of household i. Again, θ_{bh} represents a set of boundary fixed effects, equaling one if house h is within a specified distance of boundary b and zero otherwise. Each household chooses its residence h to maximize its indirect utility function V_h^i:[29]

$$(2) \quad \underset{(h)}{Max} \quad V_h^i = \alpha_X^i X_h - \alpha_p^i p_h - \alpha_d^i d_h^i + \theta_{bh} + \xi_h + \varepsilon_h^i.$$

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

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

$$(3) \quad \alpha_j^i = \alpha_{0j} + \sum_{k=1}^{K} \alpha_{kj} z_k^i,$$

with equation (3) describing household i's preference for choice characteristic j.

Given the household's problem described in equations (2)-(3), household i chooses housing choice h if the utility that it receives from this choice exceeds the utility that it receives from all other possible choices. Therefore, the probability that a household chooses any particular house depends in general on the characteristics of the full set of possible housing choices.

Estimation

---

[29] 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.

Estimation of the model follows a two-step procedure related to that in Berry, Levinsohn, and Pakes (1995).[30] It is helpful to introduce some notation to simplify the exposition. In particular, we rewrite the indirect utility function as:

$$(4) \quad V_h^i = \delta_h + \lambda_h^i + \varepsilon_h^i$$

where

$$(5) \quad \delta_h = \alpha_{0X} X_h - \alpha_{0p} p_h + \theta_{bh} + \xi_h$$

and

$$(6) \quad \lambda_h^i = \left(\sum_{k=1}^{K} \alpha_{kX} z_k^i \right) X_h - \left(\sum_{k=1}^{K} \alpha_{kp} z_k^i \right) p_h - \left(\sum_{k=1}^{K} \alpha_{kd} z_k^i \right) d_h .$$

In equation (5), δ_h captures the portion of the utility provided by housing choice h that is common to all households, and in (6), 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, namely ξ_h, captures the portion of unobserved preferences for housing choice h correlated across households, while ε_h^i represents unobserved preferences over and above this shared component.

The first step of the estimation procedure is a maximum likelihood (ML) estimator, which returns estimates of the heterogeneous parameters in λ and mean indirect utilities, δ_h. The ML estimator is based on maximizing the probability that the model correctly matches each household with its chosen housing choice. In particular, 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 h. We assume that ε_h^i is drawn from the extreme value distribution, in which case this probability can be written:

$$(7) \quad P_h^i = \exp(\delta_h + \lambda_h^i) \Big/ \sum_k \exp(\delta_k + \lambda_k^i) .$$

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

$$(8) \quad \ell = \sum_i \sum_h I_h^i \ln(P_h^i)$$

---

[30] A fuller discussion of model and estimation can be found in Bayer, McMillan, and Rueben (2004).

where I^i_h is an indicator variable that equals 1 if household i chooses housing choice h in the data and 0 otherwise. The first step of the estimation procedure then consists of searching over the parameters in λ and the vector of mean indirect utilities δ_h to maximize ℓ.

Intuitively, it is easy to see how this first step of the estimation procedure ties down the heterogeneous parameters – those involving interactions of household characteristics with housing and neighborhood characteristics. In the data, if more educated households are more likely to choose houses in neighborhoods with better schools, for instance, a positive interaction of education and average test score will allow the model to fit the data better than a negative interaction would. What is less intuitive is how the vector of mean indirect utilities is determined.

To better understand the mechanics of the first step of the estimation procedure, it is helpful to write the derivative of the log-likelihood function with respect to δ_h:

$$(9) \quad \frac{\partial \ell}{\partial \delta_h} = \sum_{i=h} \frac{\partial \ln(P^i_h)}{\partial \delta_h} + \sum_{i \neq h} \frac{\partial \ln(P^i_h)}{\partial \delta_h} = \sum_{i=h}\left(1 - P^i_h\right) + \sum_{i \neq h}\left(-P^i_h\right) = 1 - \sum_i \left(P^i_h\right) = 0.$$

As this equation shows, the likelihood function is maximized at the vector δ that forces the sum of the probabilities to equal one, $\sum_i \left(P^i_h\right) = 1$ for each house. That this condition must hold for all houses results from a fundamental trade-off in the likelihood function. In particular, an increase in any particular δ_h raises the probability that each household in the sample chooses house h. While this increases the probability that the model correctly predicts the choice of the household that actually resides in house h, it decreases the probability that all of the other households in the sample make the correct choice. In this way, the first step of the estimation approach consists of choosing the interaction parameters that best match each individual with their chosen house, while ensuring that no house is systematically more attractive than any other house, according to the metric $\sum_i \left(P^i_h\right)$. [31]

---

[31] For any set of interaction parameters (those in λ), a simple contraction mapping can be used to calculate the vector δ that solves the set of first-order conditions: $\sum_i \left(P^i_h\right) = 1, \forall\, h$. For our application, the contraction mapping is simply: $\delta^{t+1}_h = \delta^t_h - \ln(\sum_i \hat{P}^i_h)$, 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 step of the estimation procedure. 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 ML

Having estimated the vector of mean indirect utilities in the first step, the second step of the estimation approach involves decomposing δ into observable and unobservable components according to the regression equation (5). Note that equation (5), which forms the basis for the second-step regression in the estimation of the sorting model, bears more than a passing resemblance to the hedonic price regression shown in equation (1). In particular, moving price to the left-hand side of equation (5) yields:

$$(10) \qquad p_h + \frac{1}{\alpha_{0p}} \delta_h = \frac{\alpha_{0x}}{\alpha_{0p}} X_h + \frac{1}{\alpha_{0p}} \theta_{bh} + \frac{1}{\alpha_{0p}} \xi_h.$$

Consequently, 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.

A Simple Example

To provide some intuition for the relationship between the coefficients of equation (10), which provide a measure of mean preferences for each attribute, and those of equation (1), which provide a measure of the hedonic (equilibrium) price of each attribute, Figures 5 and 6 characterize a housing market equilibrium in two simple settings. Figure 5 illustrates a setting in which households value a single, discrete characteristic such as a view of the Golden Gate Bridge. In the figure, the downward-sloping line represents the marginal willingness-to-pay (MWTP) curve for the households in the market. If only a few houses in the market had a view, as represented by H_1, the hedonic price of a view would reflect the MWTP of a household with a relatively strong taste for a view, as indicated by p_1^* in the figure. If, on the other hand, a view were widely available, the price of the view would generally reflect the MWTP of someone much lower in the taste distribution, as indicated by p_2^*, for example. In this way, the equilibrium price of a view is set by the household on the margin of purchasing a house with a view, and will be a function of both its supply and the distribution of preferences.[32]

This simple example makes clear a basic feature of the relationship between hedonic prices and preferences: hedonic prices should reflect mean preferences when households are homogeneous; in this case the MWTP curve would simply be a horizontal line. This can also be seen in our model. In particular, note that when households have homogeneous preferences (up

---

estimates of the interaction parameters and the vector of mean indirect utilities δ. A detailed discussion of the asymptotic properties of δ is presented in the Technical Appendix.
[32] See Epple (1987) and Ekeland *et al.* (2004) for illuminating discussions.

to the *i.i.d.* error ε_h^i), the first-order conditions, $\sum_i \left(P_h^i \right) = 1 \ \forall \ h$, imply that the ML estimates of δ_h

must be identical (equal to a constant K) for all houses. In this case, then, equation (5) can be re-

written:

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

which is simply equation (1).[33] This equivalence makes clear that the coefficient estimates from a

hedonic price regression properly return the mean marginal valuations of housing and

neighborhood attributes when heterogeneity in preferences is limited to only an idiosyncratic

component.[34]

The Golden Gate Bridge example also provides some intuition for the way the adjustment

to the hedonic price regression in equation (10) – the mean indirect utility δ_h – is determined. In

particular, when the number of houses with a view is small (H_1 in Figure 6), the majority of

households are not willing to pay the equilibrium (hedonic) price to purchase a view. Thus, the

mean indirect utility provided by a house with a view will be less than that provided by a house

without one. In essence, the goal of the first stage of the estimation procedure is to best predict

the location decisions observed in the data. Thus in this case, in order to explain why the

majority of households choose houses without a view, the estimated values of δ_h for houses with

a view must be less than those for houses without a view. This effectively lowers the value of

houses with a view on the left hand side of equation (10), leading to an estimated mean MWTP

for a view that is lower than its hedonic price.

Of course, many housing and neighborhood characteristics are not discrete but are

supplied on a more continuous basis throughout a metropolitan area. To gain some intuition for

the relationship of the hedonic price to preferences in this case, it is helpful to consider a simple

characterization of the equilibrium when households value only a single location attribute – e.g.

school quality – that varies across the neighborhoods of the metropolitan area. Figure 6 provides

a graphical depiction of this case. Because the Bay Area contains hundreds of schools, the

equilibrium difference in housing prices between each pair of schools ranked according to quality

is the MWTP of the household on the corresponding threshold between schools. These

equilibrium prices are represent by the p_j^* terms on the vertical axis. If there are roughly an equal

---

[33] K is simply absorbed into the constant term.
[34] This condition holds no matter what assumption is made concerning the distribution of the idiosyncratic error term. Prior research by Cropper *et al.* (1993) compares hedonic and discrete-choice approaches. Unlike the current paper, their analysis looks at simulation results rather than carrying out empirical estimation, and their discrete choice model does not include unobservable choice characteristics.

number of students in each school, averaging the equilibrium price over all of the houses in the sample corresponds roughly to the mean MWTP of all households. Consequently, for attributes that vary more continuously throughout the region, there is likely to be only a slight difference between the mean preferences estimated in the heterogeneous sorting model and the coefficients of the hedonic price regression.

Forming an Instrument for Price

In addition to the vector of mean indirect utilities δ_h from the first stage of our analysis, a second piece of information is needed to estimate equation (10). When this equation is written in the way it usually appears in the IO literature (as in equation (5)), it is immediately obvious that an instrument is needed to address the likely significant correlation between housing prices and unobserved housing/neighborhood quality, ξ_h. To deal with this issue, we follow the IO literature closely by deriving a variant of the standard instrument used in the differentiated products demand literature.

The key insight from the IO literature is that the equilibrium price of any particular product will be affected not only by its own quality but also by the availability of products that are close substitutes for it. The equivalent insight in a housing market context is that two identical houses in neighborhoods of identical quality may command very different prices, depending on how they are situated within the metropolitan area. Prices might vary because of variation both in proximity to employment centers and in the quality of nearby housing alternatives. For our application, we develop an instrument for price that is based on the exogenous attributes of houses and neighborhoods that are located more than three miles away from a given house, while allowing the attributes of houses and neighborhoods within three miles of the house to directly affect utility. In this way, we assume that characteristics of houses and neighborhoods a sizeable distance away influence the equilibrium in the housing market, thereby affecting prices, but have no direct effect on utility.

To construct our instrument for price, we use a two-step procedure, beginning by estimating equation (5) with a standard set of instrumental variables. In particular, while including a full set of controls for the characteristics of the house itself and its neighborhood, as well as five variables that described land use[35] in each of the 1, 2, and 3 mile rings around the house, we instrument for price with a set of variables that describe the housing stock and land use in rings greater than three miles away. Given these initial estimates of the parameters of the

---

[35] The land use variables include percent industrial, percent commercial, percent residential, percent open space (lakes and parks), and percent other, all within given rings surrounding the house in question.

utility function, we then construct a more powerful instrument by calculating the predicted vector of market clearing prices for a version of the model that sets the vector of unobserved characteristics ξ to zero.[36] Importantly, the variation in the vector of market clearing prices over and above the variables already included as controls derives only from exogenous features of housing market in a region beyond three miles from the house in question; we use the model to concentrate this information into a single instrument that accounts for the way these features are likely to affect the equilibrium price of the house in question.

The traditional first stage of the IV estimation of equation (5) is a price regression analogous to the hedonic price regressions reported above in Table 3 but which includes the constructed instrument. In first-stage estimates for both the 0.20 and 0.10 boundary samples, the instrument enters positively and very significantly, with t-statistics of 17.7 and 10.3, respectively.

Summary of Estimation Procedure and Key Identifying Assumptions

To provide a complete picture of the assumptions maintained in our analysis, Figure 7 summarizes each step of the estimation procedure (on the left side of the figure) and highlights the corresponding assumptions needed to identify the model parameters (on the right).

If households had homogeneous preferences, then the estimation procedure would reduce to a single step, as the figure makes clear. Preferences for housing, school and neighborhood attributes could be recovered using a hedonic price regression that included boundary fixed effects. Here, the boundary fixed effects and the observed school quality would account for the correlation between neighborhood sociodemographics and unobserved neighborhood quality, while detailed controls for neighborhood sociodemographics as well as boundary fixed effects would account for the correlation between school quality and neighborhood unobservables.

In the more general case in which households have heterogeneous preferences, the first step of the estimation procedure recovers the heterogeneity parameters and the vector of mean indirect utilities by maximizing the probability that each household makes its observed housing choice, appealing to revealed preference. Regressing mean indirect utility on observables and boundary fixed effects, and instrumenting for price, the second step returns mean preferences for housing, school and neighborhood attributes.

---

[36] As shown in Bayer, McMillan, and Rueben (2004), the model developed in Section 5 can be used to characterize a sorting equilibrium with an additional assumption that prices adjust to clear the market. To construct an instrument, we simply solve for the vector of market clearing prices that corresponds to what the model would predict, given an initial estimate of the parameters and only the exogenous characteristics of houses and neighborhoods.

Heterogeneous Sorting Model

Step	Description of Estimation Procedure		Key Identifying Assumptions
1	Estimate vector of mean indirect utilities, δ, and the interaction parameters in λ in equation (4) via maximum likelihood.	1	Identification is based on the notion of revealed preference: the coefficients are selected to maximize the probability each household makes its observed housing choice.
2	Estimate IV regression of vector of mean indirect utility δ on observable characteristics and boundary fixed effects according to equation (5), using an instrumental variable for housing price.		
a	*Housing Prices* - Following IO literature, correlation between housing price and unobserved housing/neighborhood quality addressed using instrument based on exogenous characteristics of housing stock and neighborhoods beyond a 3-mile threshold.	2a	Exogenous features of housing stock and land usage located more than three miles from a house affect housing price through the market equilibrium but do not affect utility directly.
b	*School Quality* - Correlation of school quality and unobserved neighborhood quality addessed by including boundary fixed effects and detailed controls for neighborhood sociodemographics.	2b	(i) Housing characteristics vary continuously across boundaries; (ii) Measures for neighborhood race/ethnicity, education, and income included in regression control fully for sorting across boundaries.
c	*Neighborhood Sociodemographics* - Correlation between neighborhood sociodemographic composition and unobserved neighborhood quality addressed by including boundary fixed effects in the analysis.	2c	(i) Housing characteristics vary continuously across boundaries; (ii) Variation in neighborhood sociodemographics at boundaries is fundamentally driven by differences in school quality; (iii) Average test score and other school characteristics included in specifications control fully for differences in school quality.

Homogeneous Sorting Model - Hedonic Price Regression

1	Under the assumption of homogeneous preferences, estimation reduces to hedonc price regression, given by equation (1). Boundary fixed effects are included in the regression to account for endogeneity of school quality and neighborhood sociodemographics.	1	See Assumptions 2b and 2c above - the same identifying assumptions for school quality and neighborhood sociodemographics apply.

6   HETEROGENEOUS SORTING MODEL - RESULTS

Mean Preferences

The first row of Table 6 reports estimates of mean preferences for four specifications of equation (10). We focus again on results using the sample of houses within 0.20 miles of a boundary as they are more precise than the results using the sample within 0.10 miles. The estimated mean preferences for average test score are almost identical to the coefficients from the hedonic price regression. When boundary fixed effects and neighborhood sociodemographics are included in the analysis, the estimate mean MWTP for school quality is \$19.7 per month[37] compared with the estimated effect of \$17.3 on housing prices in the analogous hedonic price regression reported in the second column of Table 3. In fact, this pattern – that the coefficients in the hedonic price regression more or less captures mean preferences – holds for a number of the other housing and neighborhood characteristics that vary throughout the metropolitan area, included in the analysis but not reported here. This pattern conforms to the intuition developed in Figure 6 above.

---

[37] Though the mean direct effect of school quality on house prices estimated here appears low, we note that an increase in school quality may have an additional *indirect* effect on prices as households re-sort. (See Bayer *et al.* (2007).)

In general, when the choice problem is viewed as single-dimensional, one would expect the hedonic price regression to diverge from mean preferences only for choice characteristics that vary less continuously throughout the metropolitan region or that may be in limited supply. Notably, in our analysis, estimated mean preferences differ from the corresponding coefficient in the hedonic price regression for neighborhood race. As in the hedonic price regressions, the inclusion of boundary fixed effects substantially reduces the magnitude of the estimated mean MWTP of all of the neighborhood sociodemographic characteristics. Yet even when fixed effects are included, the estimated mean MWTP from our sorting model for black neighbors remains significantly negative, -$104 per month, and statistically significant.

That hedonic prices diverge from mean preferences in the case of neighborhood race is consistent with the notion that households can self-segregate on the basis of race without requiring any equilibrium price differences across neighborhoods. In this case, mean preferences for black neighbors would be negative because the majority of the population (around 60 percent of our boundary samples) is white, while the hedonic price regression would simply reflect the fact that a sorting equilibrium can be achieved without race being capitalized into housing prices. The estimated heterogeneity in preferences for neighborhood race is entirely consistent with this explanation; we now turn to a discussion of these heterogeneity parameters.

Heterogeneity in Preferences

Table 7 reports the implied estimates of the heterogeneity in MWTP for the average test score and neighborhood sociodemographic characteristics across households with different characteristics for our preferred specification, which includes both neighborhood sociodemographic characteristics and boundary fixed effects.[38]

The estimates of the heterogeneity in the MWTP for neighborhood sociodemographic characteristics reveal a fascinating asymmetry: while all households prefer to live in higher-income neighborhoods, *conditional on neighborhood income* households prefer to self-segregate on the basis of both race and education. In particular, the estimates imply that college-educated households are willing to pay $58 per month more than those without a college degree to live in a neighborhood that has 10 percent more college-educated households. When combined with the estimated mean MWTP of $10 per month reported in the first row, this estimate implies that households at each level of educational attainment prefer neighbors with like education levels:

---

[38] The full heterogeneous choice model includes 135 interactions between nine household characteristics and fifteen housing and neighborhood characteristics. In Table 7 we only report MWTP for test scores and sociodemographics which correspond to the core of our analysis. The full set of included variables is listed in the note to Table 7.

while college-educated households would pay an addition $32 per month to live in a neighborhood that had 10 percent more college-educated households, households without a college degree would actually need *compensating* to live in a neighborhood with 10 percent more college-educated neighbors, to the tune of $26 per month. Note that the preference for self-segregation on the basis of educational attainment is somewhat stronger for college-educated households.

Similarly, the heterogeneity estimates imply that blacks are willing to pay $98 more per month than whites to live in a neighborhood that has 10 percent more black versus white households. The mean MWTP for such an increase is -$10.5 per month, primarily reflecting the negative valuation of the white majority. Thus $98 is the difference between the *positive* MWTP of black households for this change and the *negative* MWTP of white households, indicating that households have strong self-segregating racial preferences.[39]

Focusing on the heterogeneity in tastes for school quality, a household's willingness-to-pay increases with income, the presence of children, education, employment, and age. Blacks have a significantly lower willingness to pay for school quality relative to whites, although this may be related to unobservable factors such as the substantial degree of wealth inequality across race. The presence of children increases demand for school quality. That it does not increase demand by a greater amount may reflect the fact that the presence of children also raises the desired levels of other forms of consumption. The parameter estimates not presented in the table, for example, reveal that households with children have a much greater demand for larger houses.

As one might expect, increases in household income and education (which may proxy better for lifetime income) are associated with increased demand for better schools. They are also associated with higher demand for more educated and higher-income neighbors. We discuss possible consequences of this configuration of preferences in the next subsection.

Discussion

Taken together, the estimates of the heterogeneous model of sorting reveal a number of key findings. First, the estimated mean preferences for housing and neighborhood characteristics that vary more or less continuously throughout the metropolitan area closely resemble the estimates of a simple hedonic price regression. This suggests that the estimated coefficients for

---

[39] It is also important to point out that these interactions pick up any direct preferences for living near others of the same race (e.g., a recent immigrant from China may want to interact with neighbors who also have immigrated from China) as well as any unobservable neighborhood or housing amenities valued more strongly by households of this group (e.g., recent immigrants from China may have similar tastes for shops, restaurants, and other neighborhood amenities).

these types of variable in a hedonic price regression may generally be interpreted not only as a measure of the implicit price of a particular attribute in the housing market but also as a reasonable estimate of mean preferences. This additional interpretation of some of the coefficients from hedonic price regressions is reassuring given that it is generally difficult to obtain the kind of data necessary to estimate the heterogeneous model presented here – i.e., data that precisely match households to their houses and neighborhoods.

The estimates of the heterogeneous model of sorting along with the hedonic price regression results reported in Section 3 tell a coherent story regarding the role of race in the housing market. In particular, they suggest that (i) neighborhood race is strongly correlated with unobserved housing and neighborhood quality, (ii) households have strong self-segregating preferences, and (iii) neighborhood race may not be directly capitalized into housing prices as neighborhood price differences are not required to clear the market.

The results also reveal a similar pattern for neighborhood education, implying both that households prefer to self-segregate on the basis of education and that the average education of a neighborhood tends to be highly correlated with unobserved neighborhood quality. Taken together, however, the results tell a very different story for neighborhood income, implying that all households place significant value on richer neighbors.

Finally, the particular combination of heterogeneous preferences for school quality (with better-educated and higher-income households having higher demands) and heterogeneous preferences for neighbors (with better-educated households having strong preferences for living with highly educated neighbors) suggests that exogenous changes in school quality may have compounding general equilibrium effects. In particular, an exogenous change in school quality would be likely to have both a direct effect on housing prices associated with preferences for higher school quality in addition to indirect effects, as households re-sorted. Our estimates suggest that the improvement in a given school's quality would disproportionately attract more highly educated households to the neighborhood, in turn making the neighborhood even more attractive to higher-income, highly educated households, and raising house prices further. Such second-round 'social multiplier' effects on prices could potentially be greater than the direct effect.[40]

---

[40] The preference estimates from the current analysis provide an important input when examining these issues. In related work (see Bayer, Ferreira and McMillan (2007)), we use general equilibrium simulations based on the estimates reported in this paper to explore the size of social multiplier effects associated with increases in school quality as households re-sort.

7   CONCLUSION

Household sorting induces correlations among observed and unobserved neighborhood attributes, making it difficult to infer the nature of the preferences that drive the sorting process. Given the scarcity of research designs that deal effectively with the resulting endogeneity problem, the boundary discontinuity design (BDD) has attracted widespread attention, providing a straightforward way to estimate the value of amenities (such as school quality) that vary discontinuously across well-defined boundaries.

Yet sorting has several implications for the use of the boundary discontinuity approach, as we have argued. First, discontinuous local amenities are likely to generate sorting with respect to the boundary, so neighborhood sociodemographics also vary discontinuously there. This implies that any house price differences across boundaries are likely to overstate the value of the discontinuous local amenity, and that better estimates can be achieved by controlling carefully for the characteristics of immediate neighbors. Second, to the extent that researchers can control for the fundamental source of the sorting at the boundary – in our case, differences in school quality – so any variation in neighborhood sociodemographics across boundaries is likely to be close-to-uncorrelated with unobserved housing and neighborhood attributes. Thus, a BDD provides a reasonable way to address the challenging endogeneity of neighborhood sociodemographics.

Sorting also naturally indicates that households are heterogeneous in their willingness to pay for housing and neighborhood attributes. At the heart of the analysis, we develop a heterogeneous sorting model that embeds a BDD, showing how this approach can be used to identify the full distribution of household preferences for housing and neighborhood attributes. Taking advantage of unusually rich data from the Bay Area, the analysis shows clearly that households sort with respect to school attendance zone boundaries, and that OLS estimates of the capitalization of neighborhood sociodemographics into housing prices are significantly overstated, due to the correlation of these characteristics with unobserved neighborhood quality. Conditional on income, the results also imply that households prefer to self-segregate on the basis of education and especially race.

This sorting model provides a natural device for exploring the general equilibrium implications of the preference estimates, using counterfactual simulations. In an education context, these would complement recent research that has used calibrated equilibrium models to simulate policy changes, uncovering interesting general equilibrium effects in the process – see Epple and Romano (1998), Nechyba (1999, 2000), and Fernandez and Rogerson (2003). An appealing feature of the current framework is that it permits the direct estimation of a broad range

29

of preference parameters influencing the sorting process, with the potential to improve our understanding of policy reforms and the workings of the urban economy more widely.

REFERENCES

Bajari, Patrick and Lanier Benkard (2005), "Demand Estimation with Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach," *Journal of Political Economy*, 113(6): 1239-1276.

Bajari, Patrick, and Matthew Kahn (2005), "Estimating Housing Demand with an Application to Explaining Racial Segregation in Cities," *Journal of Business & Economic Statistics*, American Statistical Association, 23: 20-33.

Barrow, Lisa (2002), "School Choice through Relocation: Evidence from the Washington, D.C. area," *Journal of Public Economics*, 86(2): 155-189.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan (2007), "Tiebout Sorting, Social Multipliers, and the Demand for School Quality," mimeo.

Bayer, Patrick, Robert McMillan, and Kim Rueben (2004), "An Equilibrium Model of Sorting in an Urban Housing Market," NBER Working Paper 10865.

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

Berry, Steven, Oliver Linton, and Ariel Pakes (2004), "Limit Theorems for Estimating the Parameters of Differentiated Product Demand Systems," *Review of Economic Studies*.

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

Cook, Thomas, and Donald Campbell (1979), *Quasi-experimentation: Design and Analysis Issues for Field Settings*, Boston: Houghton Mifflin.

Cropper, Maureen, Leland Deck, Nalin Kishor, and Kenneth McConnell (1993), "Valuing Product Attributes Using Single Market Data: A Comparison of Hedonic and Discrete Choice Approaches," *Review of Economics and Statistics*, 75(2): 225-232.

Cutler, David, Edward Glaeser, and Jacob Vigdor (1999), "The Rise and Decline of the American Ghetto," *Journal of Political Economy*, 107: 455-506.

Ekeland, Ivar, James Heckman, and Lars Nesheim (2004), "Identification and Estimation of Hedonic Models," *Journal of Political Economy*, 112: 60-109.

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*, Vol. 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*, Vol. 23, pp. 585-610.

Epple, Dennis and Richard Romano (1998), "Competition Between Private and Public Schools, Vouchers, and Peer Group Effects," *American Economic Review* 88(1): 33-62.

Epple, Dennis and Allan Zelenitz (1981), "The Implications of Competing among Jurisdictions: Does Tiebout Need Politics?," *Journal of Political Economy*, 89: 1197-1217.

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 (2003), "Equity and Resources: An Analysis of Educational Finance Systems," *Journal of Political Economy*, vol. 111, no. 4.

Heckman, James, Rosa Matzkin, and Lars Nesheim (2003), "Simulation and Estimation of Hedonic Models," unpublished manuscript, University of Chicago.

Kane, Thomas, Douglas Staiger, and Gavin Samms (2003), "School Accountability Ratings and Houise Values," *Brookings-Wharton Papers on Urban Affairs*.

Lee, David (2007), "Randomized Experiments from Non-random Selection in U.S. House Elections," *Journal of Econometrics,* forthcoming.

McFadden, Daniel (1973), "Conditional Logit Analysis of Qualitative Choice Behavior," *Frontiers in Econometrics*, Academic Press: New York, p. 105-142.

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. (1997), "Existence of Equilibrium and Stratification in Local and Hierarchical Tiebout Economies with Property Taxes and Voting," *Economic Theory*, Vol. 10, pp. 277-304.

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* 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, pp. 51-73.

Nesheim, Lars (2001), "Equilibrium Sorting of Heterogeneous Consumers Across Locations: Theory and Empirical Implications," Ph. D. Dissertation, University of Chicago.

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.

Rothstein, Jesse (2006), "Good Principals or Good Peers: Parental Valuation of School Characteristics, Tiebout Equilibrium, and the Incentive Effects of Competition among Jurisdictions," *American Economic Review*, 96(4): 1333-1350.

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

Figure 1: Test scores and house prices around the boundary
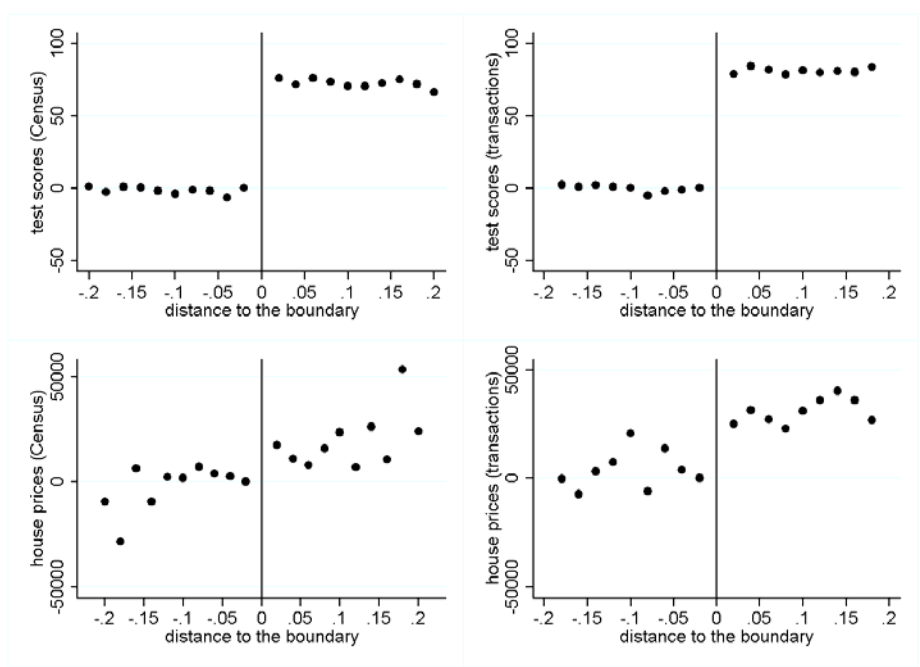
Notes: Each panel in this figure is constructed using the following procedure: (i) regress the variable in question on boundary fixed effects and on 0.02 mile band distance-to-the-boundary dummy variables; (ii) plot the coefficients on these distance dummies. Thus a given point in each figure represents this conditional average at a given distance to the boundary, where negative distances indicate the 'low' test score side.

Figure 2: Census housing characteristics around the boundary
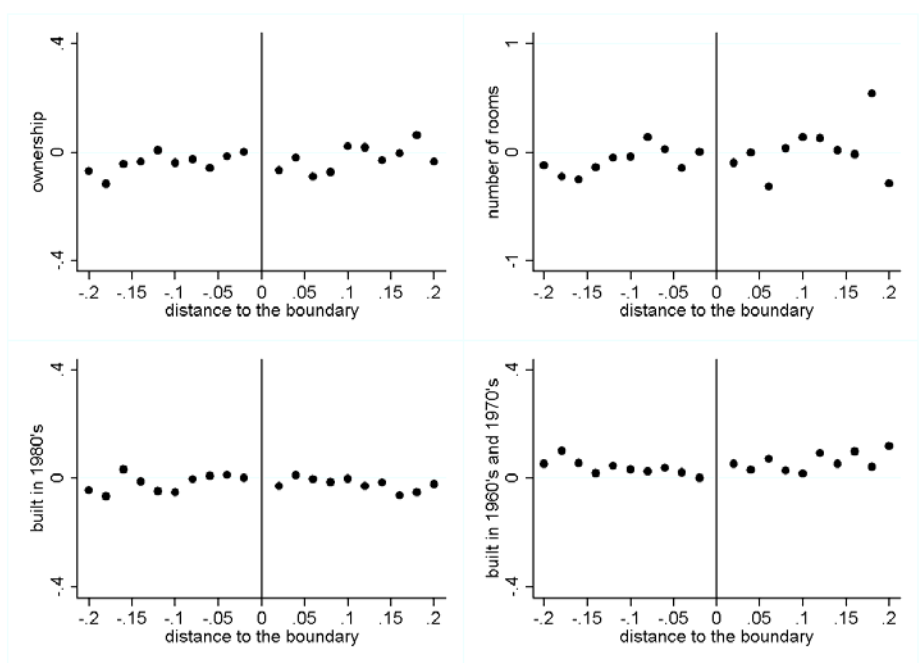
Notes: Each panel in this figure is constructed using the following procedure: (i) regress the variable in question on boundary fixed effects and on 0.02 mile band distance-to-the-boundary dummy variables; (ii) plot the coefficients on these distance dummies. Thus a given point in each figure represents this conditional average at a given distance to the boundary, where negative distances indicate the 'low' test score side.

Figure 3: Transactions data housing characteristics around the boundary
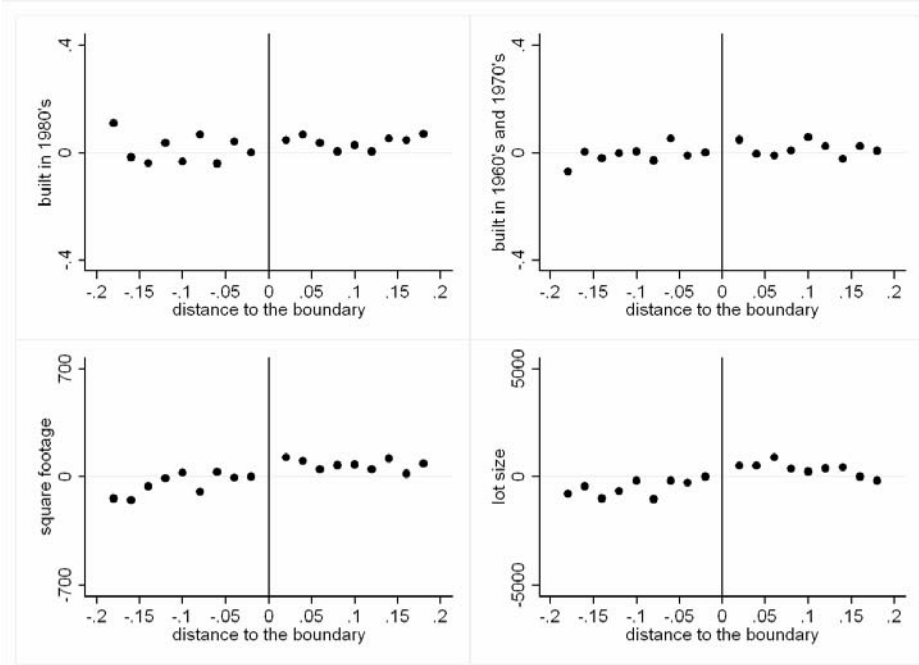
Notes: Each panel in this figure is constructed using the following procedure: (i) regress the variable in question on boundary fixed effects and on 0.02 mile band distance-to-the-boundary dummy variables; (ii) plot the coefficients on these distance dummies. Thus a given point in each figure represents this conditional average at a given distance to the boundary, where negative distances indicate the 'low' test score side.

Figure 4: Neighborhood sociodemographics around the boundary
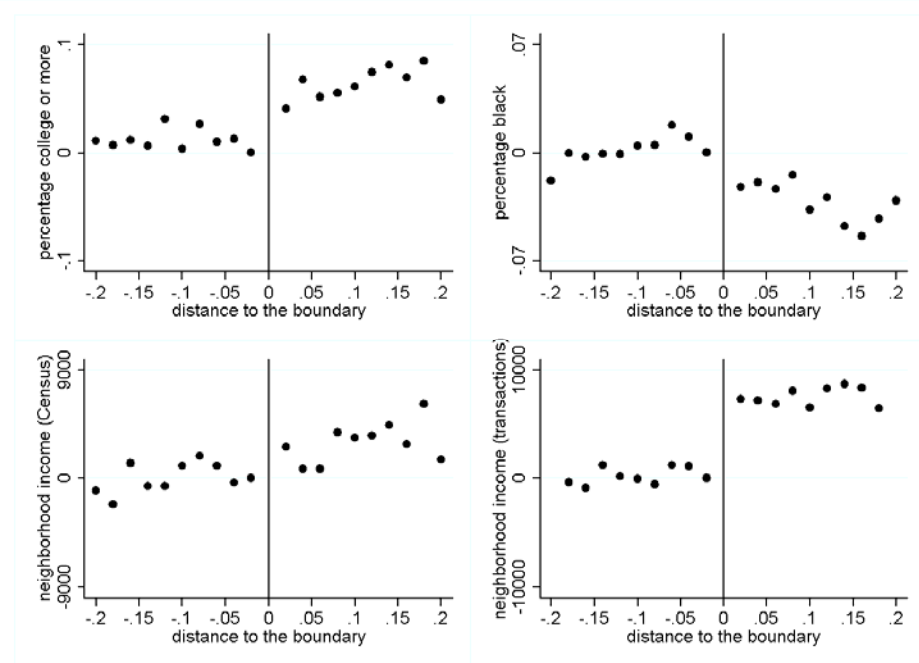
Notes: Each panel in this figure is constructed using the following procedure: (i) regress the variable in question on boundary fixed effects and on 0.02 mile band distance-to-the-boundary dummy variables; (ii) plot the coefficients on these distance dummies. Thus a given point in each figure represents this conditional average at a given distance to the boundary, where negative distances indicate the 'low' test score side.

Figure 5: Demand for a View of the Golden Gate Bridge

Price ($)

p_1^*

MWTP Curve   Mean MWTP

p_M^*

p_2^*

H_1 H_2 Number of Houses
With a View
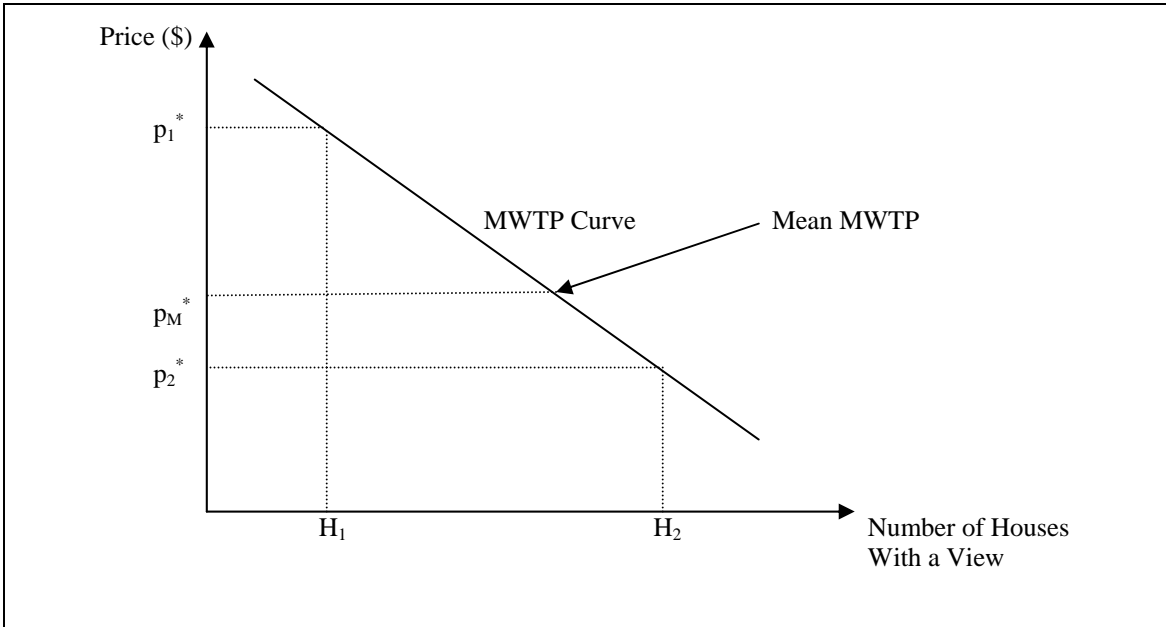
Figure 6: Demand for School Quality

Price ($)

p_1^*

p_2^*

p_3^* MWTP Curve   Mean MWTP

p_4^*

p_5^*

p_6^*

S_1 S_2 S_3 S_4 S_5 S_6 S_7 Allocation of Houses
To Schools With
Quality $\{S_1, \ldots S_7\}$

Table 1. Sample Statistics Comparing the Full Sample with Houses within 0.20 miles of a Boundary

Sample	full sample		within 0.20 miles of boundaries				
			boundary sample	high test score side	low test score side	difference in means	test of difference
Observations	242,100		27,548	13,612	13,936		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean	S.D.	Mean	Mean	Mean	((4) - (5))	t-statistic
Housing Prices							
House value (if owned)	297,700	178,479	250,005	259,475	240,756	18,719	4.15
Monthly rent (if rented)	744	316	678	688	669	18.80	1.73
School Quality							
Average test score	527	74	507	544	471	74	25.44
Housing Characteristics							
1 if unit owned	0.60	0.49	0.54	0.55	0.53	0.02	0.89
Number of rooms	5.11	1.99	4.96	5.02	4.90	0.12	1.56
1 if built in 1980s	0.14	0.35	0.11	0.11	0.11	0.00	-0.31
1 if built in 1960s or 1970s	0.39	0.49	0.34	0.35	0.33	0.01	0.84
Elevation	210	179	176	178	173	6	1.64
Population density	0.43	0.50	0.39	0.38	0.40	-0.02	-1.38
Neighborhood Sociodemographics							
% Census block group white	0.68	0.23	0.61	0.63	0.60	0.03	3.40
% Census block group black	0.08	0.16	0.18	0.17	0.20	-0.03	-3.15
% Census block group coll deg or more	0.44	0.20	0.41	0.44	0.39	0.05	6.18
Average block group income	54,742	26,075	46,271	47,718	44,857	2,861	2.61

Note: This table reports summary statistics for the key variables included in the analysis. The boundary sample includes all houses located within 0.20 miles of a boundary with another school attendance zone. A house is considered to be on the 'high' ('low') side of a boundary if the test score at its local school is greater (less) than the corresponding test score for the closest house on the opposite side of an attendance zone boundary. Sample statistics are reported for the high- and low-side of boundaries for which the test score gap is in excess of the median gap (38.4 points) in columns (4) and (5), respectively. Column (7) reports the t-statistic for a test of the hypothesis that the mean of the variable listed in the row heading does not vary across school attendance zone boundaries. This test conditions on boundary fixed effects (so as to compare houses on opposite sides of the same boundary) and adjusts for the clustering of observations at the Census block group level.
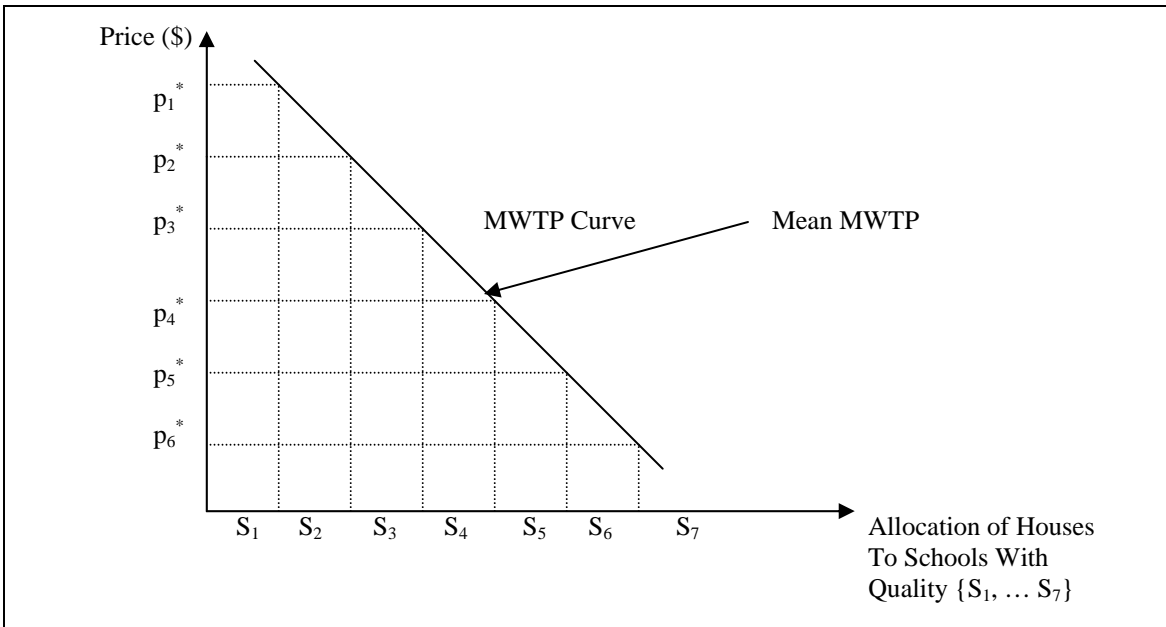
Table 2. Sample Statistics Comparing the Full Sample with Houses within 0.10 miles of a Boundary

Sample	full sample		within 0.10 miles of boundaries				
			boundary sample	high test score side	low test score side	difference in means	test of difference
Observations	242,100		15,122	7,824	7,298		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean	S.D.	Mean	Mean	Mean	((4) - (5))	t-statistic
Housing Prices							
House value (if owned)	297,700	178,479	244,506	251,742	236,749	14,993	3.95
Monthly rent (if rented)	744	316	667	673	661	11.8	1.05
School Quality							
Average test score	527	74	505	542	466	75	21.00
Housing Characteristics							
1 if unit owned	0.60	0.49	0.51	0.51	0.51	0.00	-0.20
Number of rooms	5.11	1.99	4.88	4.88	4.87	0.01	0.14
1 if built in 1980s	0.14	0.35	0.12	0.12	0.11	0.01	1.12
1 if built in 1960s or 1970s	0.39	0.49	0.30	0.30	0.30	0.01	0.48
Elevation	210	179	164	166	162	4	1.53
Population density	0.43	0.50	0.41	0.41	0.41	0.00	-0.20
Neighborhood Sociodemographics							
% Census block group white	0.68	0.23	0.60	0.61	0.58	0.03	2.82
% Census block group black	0.08	0.16	0.20	0.19	0.22	-0.03	-3.13
% Census block group coll deg or more	0.44	0.20	0.41	0.43	0.38	0.05	5.24
Average block group income	54,742	26,075	44,831	45,657	43,945	1,711	1.70

Note: This table reports summary statistics for the key variables included in the analysis. The boundary sample includes all houses located within 0.20 miles of a boundary with another school attendance zone. A house is considered to be on the 'high' ('low') side of a boundary if the test score at its local school is greater (less) than the corresponding test score for the closest house on the opposite side of an attendance zone boundary. Sample statistics are reported for the high- and low-side of boundaries for which the test score gap is in excess of the median gap (38.4 points) in columns (4) and (5), respectively. Column (7) reports the t-statistic for a test of the hypothesis that the mean of the variable listed in the row heading does not vary across school attendance zone boundaries. This test conditions on boundary fixed effects (so as to compare houses on opposite sides of the same boundary) and adjusts for the clustering of observations at the Census block group level.

Table 3: Key Coefficients from Baseline Hedonic Price Regressions

Sample Observations Boundary Fixed Effects	Within 0.20 Miles of Boundary 27,548		Within 0.10 Miles of Boundary 15,122	
	No	Yes	No	Yes
Panel A: Excluding Neighborhood Sociodemographic Characteristics	**(1)**	**(2)**	**(5)**	**(6)**
average test score (in standard deviations)	123.7	33.1	126.5	26.1
	(13.2)	(7.6)	(12.4)	(6.6)
R^2	0.54	0.62	0.54	0.62
Panel B: Including Neighborhood Sociodemographic Characteristics	**(3)**	**(4)**	**(7)**	**(8)**
average test score (in standard deviations)	34.8	17.3	44.1	14.6
	(8.1)	(5.9)	(8.5)	(6.3)
% Census block group black	-99.8	1.5	-123.1	4.3
	(33.4)	(38.9)	(32.5)	(39.1)
% block group college degree or more	220.1	89.9	204.4	80.8
	(39.9)	(32.3)	(40.8)	(39.7)
average block group income (/10000)	60.0	45.0	55.6	42.9
	(4.0)	(4.6)	(4.3)	(6.1)
R^2	0.59	0.64	0.59	0.63

Note: All regressions shown in the table also include controls for whether the house is owner-occupied, the number of rooms, year built (1980s, 1960-1979, pre-1960), elevation, population density, crime, land use (% industrial, % residential, % commercial, % open space, % other) in 1, 2, and 3 mile rings around each location. The dependent variable is the monthly user cost of housing, which equals monthly rent for renter-occupied units and a monthly user cost for owner-occupied housing, calculated as described in the text. Standard errors corrected for clustering at the school level are reported in parentheses.

Table 4: Hedonic Price Regressions - Average Test Score: Alternative Samples

Sample	Within 0.20 miles of Boundary			
Neighborhood Sociodemographics	Excluded		Included	
Boundary Fixed Effects	No	Yes	No	Yes
Coefficient on Average Test Score (standard deviation)	**(1)**	**(2)**	**(3)**	**(4)**
Baseline Results	123.7	33.1	34.8	17.3
N = 27,548	(13.2)	(7.6)	(8.1)	(5.9)
Schools versus Immediate Neighbors				
(A) Including School Peer and Teacher Measures	95.0	32.1	31.5	22.6
N = 27,548	(17.9)	(10.4)	(9.3)	(8.5)
Alternative Measures of Neighborhood Characteristics				
(B) Including Block and Block Group Measures			36.0	19.8
N = 27,548			(7.8)	(5.7)
(C) Including Block and Alternative Block Group Measure			33.7	23.8
N = 27,548			(7.3)	(5.6)
Other Robustness Checks				
(D) Dropping Top-Coded Houses	86.6	29.5	20.3	16.1
N = 26,579	(9.9)	(6.6)	(7.7)	(5.7)
Only Owner-Occupied Housing Units				
(E) Using Census Reported House Value	64,891	14,874	27,883	9,376
N = 15,139	(7,474)	(3,197)	(5,047)	(2,460)
(F) Using Prices from Transactions Sample	34,262	12,210	14,208	9,176
N = 10,171	(4,958)	(3,108)	(2,886)	(2,738)

Note: The dependent variable in specifications (A)-(D) is the monthly user cost of housing, which equals monthly rent for renter-occupied units and a monthly user cost for owner-occupied housing, calculated as described in the text; the dependent variable in specification (E) is the market value of the house self-reported in the Census; the dependent variable in specification (F) is the transaction price reported in our transactions dataset. Specifications (A)-(E) are based on our Census sample and include controls for whether the house is owner-occupied, the number of rooms, year built (1980s, 1960-1979, pre-1960), elevation, population density, crime, land use (% industrial, % residential, % commercial, % open space, % other) in 1, 2 and 3 mile rings around each location. Specification (F) is based on our transactions dataset and includes the same controls as in the other specifications along with additional controls for square footage and lot size. Standard errors corrected for clustering at the school level are reported in parentheses.

Table 5: Hedonic Price Regressions - Key Neighborhood Sociodemographic Characteristics: Alternative Samples

| Sample | Within 0.20 miles of Boundary | | | | | |
| Boundary Fixed Effects | No | | | Yes | | |
	% Black	% College-Ed.	Avg Income (/$10K)	% Black	% College-Ed.	Avg Income (/$10K)
Baseline Results	-99.8	220.1	60.0	1.5	89.9	45.0
N = 27,548	(33.4)	(39.9)	(4.0)	(38.9)	(32.3)	(4.6)
Schools versus Immediate Neighbors						
(A) Including School Peer and Teacher Measures	-38.2	215.6	60.7	-13.4	97.8	45.4
N = 27,548	(36.2)	(41.0)	(4.0)	(40.3)	(32.9)	(4.6)
Alternative Measures of Neighborhood Characteristics						
(B) Including Block and Block Group Measures						
Block Measures	2.9	63.2	28.4	10.3	58.8	24.9
N = 27,548	(17.6)	(13.3)	(1.9)	(17.0)	(11.2)	(1.6)
Block Group Measures	-98.0	145.1	37.3	-13.7	36.0	25.2
N = 27,548	(37.6)	(42.2)	(4.1)	(38.8)	(34.0)	(4.4)
(C) Including Block and Alternative Block Group Measure						
Block Measures	0.0	65.6	29.7	8.2	63.7	26.3
N = 27,548	(17.2)	(12.7)	(2.0)	(16.8)	(10.9)	(1.7)
Block Group Measures	-101.2	126.4	42.2	-4.2	-12.5	30.3
N = 27,548	(37.1)	(43.5)	(4.1)	(45.3)	(41.4)	(4.2)
Other Robustness Checks						
(D) Dropping Top-Coded Houses	-116.6	229.9	47.0	1.3	129.2	38.0
N = 26,579	(31.0)	(39.6)	(5.4)	(38.9)	(34.1)	(5.2)
Only Owner-Occupied Housing Units						
(E) Using Census Reported House Value	-54,289	46,071	25,816	-4,101	-12,437	14,353
N = 15,139	(16,013)	(21,796)	(1,997)	(12,407)	(15,899)	(1,864)
(F) Using Prices from Transactions Sample			15,810			6,780
N = 10,171			(2,470)			(1,990)

Note: The dependent variable in specifications (A)-(D) is the monthly user cost of housing, which equals monthly rent for renter-occupied units and a monthly user cost for owner-occupied housing, calculated as described in the text; the dependent variable in specification (E) is the market value of the house self-reported in the Census; the dependent variable in specification (F) is the transaction price reported in our transactions dataset. Specifications (A)-(E) are based on our Census sample and include controls for whether the house is owner-occupied, the number of rooms, year built (1980s, 1960-1979, pre-1960), elevation, population density, crime, land use (% industrial, % residential, % commercial, % open space, % other) in 1, 2 and 3 mile rings around each location. Specification (F) is based on our transactions dataset and includes the same controls as in the other specifications along with additional controls for square footage and lot size. Standard errors corrected for clustering at the school level are reported in parentheses.

Table 6: Delta Regressions - Implied Mean Willingness to Pay

Sample	Within 0.20 Miles of Boundary	
Observations	27,458	
Boundary Fixed Effects	No	Yes
Panel A: Excluding Neighborhood Sociodemographic Characteristics	**(1)**	**(2)**
average test score (in standard deviations)	97.3	40.8
	(14.0)	(5.5)
Panel B: Including Neighborhood Sociodemographic Characteristics	(3)	(4)
average test score (in standard deviations)	18.0	19.7
	(8.3)	(7.4)
% block group black	-404.8	-104.8
	(41.4)	(36.9)
% block group college degree or more	183.5	104.6
	(26.4)	(31.8)
average block group income (/10000)	30.7	36.3
	(3.7)	(6.6)

Note: All regressions shown in the table also include controls for whether the house is owner-occupied, the number of rooms, year built (1980s, 1960-1979, pre-1960), elevation, population density, crime, land use (% industrial, % residential, % commercial, % open space, % other) in 1, 2 and 3 mile rings around each location. The dependent variable is the monthly user cost of housing, which equals monthly rent for renter-occupied units and a monthly user cost for owner-occupied housing, calculated as described in the text. Standard errors corrected for clustering at the school level are reported in parentheses.

Table 7. Heterogeneity in Marginal Willingness to Pay for Average Test Score and Neighborhood Sociodemographic Characteristics

| | Average Test Score +1 s.d. | Neighborhood Sociodemographics | | |
		+10% Black vs. White	+10% College Educated	Blk Group Avg Income + $10,000
Mean MWTP	19.69	-10.50	10.46	36.3
	(7.41)	(3.69)	(3.18)	(6.60)
Household Income (+$10,000)	1.38	-1.23	1.41	0.86
	(0.33)	(0.37)	(0.21)	(0.12)
Children Under 18 vs.	7.41	11.86	-16.07	2.37
No Children	(3.58)	(3.03)	(2.25)	(1.17)
Black vs. White	-14.31	98.34	18.45	-1.16
	(7.36)	(3.93)	(4.52)	(2.24)
College Degree or More vs.	13.03	9.19	58.05	0.31
Some College or Less	(3.57)	(3.14)	(2.33)	(1.40)

Note: The first row of the table reports the mean marginal willingness-to-pay for the change reported in the column heading. The remaining rows report the difference in willingness to pay associated with the change listed in the row heading, holding all other factors equal. The full heterogeneous choice model includes 135 interactions between nine household characteristics and fifteen housing and neighborhood characteristics. The included household characteristics are household income, the presence of children under 18, and the race/ethnicity (Asian, black, Hispanic, white), educational attainment (some college, college degree or more), work status, and age of the household head. The housing and neighborhood characteristics are the monthly user cost of housing, distance to work, average test score, whether the house is owner-occupied, number of rooms, year built (1980s, 1960-1979, pre-1960), elevation, population density, crime, and the racial composition (% Asian, % black, % Hispanic, % white) and average education (% college degree) and household income for the corresponding Census block group. Standard errors are reported in parentheses.

Appendix Table 1. Sample Statistics for Key Variables in Transactions Dataset

Sample	full sample		within 0.20 miles of boundaries			within 0.10 miles of boundaries		
			boundary sample	difference in means		boundary sample	difference in means	
Observations	266,996		10,171			4,805		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mean	S.D.	Mean	difference	t-statistic	Mean	difference	t-statistic
Housing/Neighborhood Characteristics								
Housing Prices								
transaction price	253,498	195,996	197,254	24,772	4.38	191,573	21,355	4.04
School Quality								
average test score	573	93	557	82	20.99	550	84	18.61
Housing Characteristics								
number of rooms	6.70	1.97	6.26	0.31	2.98	6.29	0.18	1.55
1 if built in 1980s	0.30	0.46	0.20	0.03	1.09	0.18	0.03	1.08
1 if built in 1960s or 1970s	0.31	0.46	0.23	0.02	0.90	0.23	0.01	0.24
square footage	1,653	712	1,483	111	3.16	1,471	94	2.83
lotsize	7,199	7,654	6,583	833	1.83	6,527	758	1.28
Neighborhood Sociodemographics (calculated within sample)								
average block group income	96,099	17,649	75,877	6,845	3.91	74,712	4,728	2.40

Note: This table reports summary statistics for the key variables in our transactions dataset. As in the Census dataset, a house is considered to be on the 'high' ('low') side of a boundary if the test score at its local school is greater (less) than the corresponding test score for the closest house on the opposite side of an attendance zone boundary. Sample statistics are reported for the full sample in columns (1) and (2). For the 0.20-mile sample, the sample mean, difference in means on the high- versus low-side of boundaries for which the test score gap is in excess of the median gap (38.4 points), and the t-statistic for a test of the hypothesis that the mean does not vary across school attendance zone boundaries are reported in columns (3)-(5), respectively. As in Tables 1 and 2, the test conditions on boundary fixed effects and adjusts for the clustering of observations at the Census block group level. Columns (6)-(8) report analogous numbers for the 0.10-mile boundary sample.

Appendix Table 2: Comparing Hedonic Price Coefficients in Full and Boundary Samples

Sample	Full Sample	Within 0.20 Miles of Boundary	Within 0.10 Miles of Boundary
Observations	242,100	27,548	15,122
Boundary Fixed Effects	No	No	No
Panel A: Excluding Neighborhood Sociodemographic Characteristics	**(1)**	**(3)**	**(5)**
average test score (in standard deviations)	129.6	123.7	126.5
	(8.8)	(13.2)	(12.4)
R^2	0.50	0.54	0.54
Panel B: Including Neighborhood Sociodemographic Characteristics	**(2)**	**(4)**	**(6)**
average test score (in standard deviations)	35.3	34.8	44.1
	(6.5)	(8.1)	(8.5)
% census block group black	-183.0	-99.8	-123.1
	(24.9)	(33.4)	(32.5)
% block group college degree or more	281.8	220.1	204.4
	(42.1)	(39.9)	(40.8)
average block group income (/10000)	64.5	60.0	55.6
	(3.0)	(4.0)	(4.3)
R^2	0.59	0.59	0.59

Note: All regressions shown in the table also include controls for whether the house is owner-occupied, the number of rooms, year built (1980s, 1960-1979, pre-1960), elevation, population density, crime, land use (% industrial, % residential, % commercial, % open space, % other) in 1, 2 and 3 mile rings around each location. The dependent variable is the monthly user cost of housing, which equals monthly rent for renter-occupied units and a monthly user cost for owner-occupied housing, calculated as described in the text. Standard errors corrected for clustering at the school level are reported in parentheses.

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.