# *Isolating the School Quality Premium in Housing Markets**

Arman Khachiyan
Thesis Advisor: Professor Jesse Rothstein

*University of California, Berkeley*
*Department of Economics*

May 2013

**Abstract**

This paper uses a pairing process to test the hypothesis that home prices are affected by school district quality. Combining a hedonic price model with a regression discontinuity design, I compare the selling prices of pairs of homes that are no more than 0.4 miles from each other and are highly similar across all other characteristics, but which lie in school districts of different quality. My unique pairing methodology more robustly controls for the confounding effects of home characteristics than previous research which relied on linear regression models. I estimate that a one standard deviation increase in school district quality is associated with a 3.3% increase in home prices.

* I would like to thank Jesse Rothstein for his ongoing guidance and support, Amar Mann and Tian Luo for their contributions in the geographic coding process, and Marina Guevorkian for access to an extensive home sales database. All mistakes are my own.

# Table of Contents

# I. Introduction

As the father of American Public Education, Horace Mann envisioned education as a birthright for American children, available and equal for all.[1] His belief in public education as our society's "Great Equalizer" has since become a cornerstone of American culture. This culture is particularly manifested in economic philosophy that assumes the equality of opportunity. Despite this philosophy, the mechanisms through which public education is organized and implemented in the United States exhibit serious and restrictive barriers to equality of access.

A direct economic approach to quantifying the inequality of access in public education is to show that home prices are higher within the attendance boundaries of high performing schools. By showing that there is a dollar value—nested in real estate markets—that grants access to the best school districts, this field of study exposes a quantifiable avenue through which quality of education is allocated according to family wealth.

As identified by Tiebout in his pivotal 1956 paper, neighborhood characteristics such as school quality drive individuals and families to "vote with their feet" when selecting a neighborhood in which to live[2]. Incorporating a scarcity problem to Tiebout's model, this sorting creates a marketplace for neighborhood services. Neighborhoods with the most valued services will, ceteris peribus, become more expensive to live in.

A hedonic price model is ideal for quantifying the value of school quality as a neighborhood service because it specifies that consumers separately value each

[1] The Republic and the School, Horace Mann, 1957
[2] A Pure Theory of Local Expenditures, Charles Tiebout, 1956

characteristic (both individual and neighborhood) of a home[3]. In this model a buyer's demand for a home is an aggregation of their demand for each individual attribute of the home. In a direct comparison between two homes in exactly the same location and with all other attributes being equal, the hedonic model holds that if school quality is a valued attribute of a home, then the price of the home in the better school district will reflect that value for schooling.

Of course, there are no two homes with the exact same location, all of the same home characteristics, and different school districts. To address this issue, the modern literature in the field has focused on a boundary discontinuity design. By comparing homes that are within a narrow range of school district boundaries, these studies mitigate the effects of neighborhood differences.

Sandra Black was the first to apply a boundary discontinuity design towards quantifying the school quality premium in housing markets[4]. She replaced the standard vector of neighborhood characteristics with a vector of boundary dummies, effectively comparing the mean selling price of homes on either side of each attendance boundary, while controlling for home characteristics.

A more recent study focused on the San Francisco Bay Area (Bayer, Ferreira, McMillan, 2007) compares census blocks on opposite sides of elementary school boundaries, while taking into consideration the contribution of socio-economic and racial sorting trends to differences in home prices[5]. Adding to the work of Sandra Black, their study similarly relies on a linear regression model to control for housing characteristics, with the specific improvement of a wide range of neighborhood characteristics available

[3] Valuing Product Attributes Using Single Market Data, Cropper, Deck, Kishor, McConnel, 1993
[4] Do Better Schools Matter? Parental Valuation of Elementary Education, Sandra Black, 1999
[5] A Unified Framework for Measuring Preferences for Schools and Neighborhoods, Bayer, Ferreira, McMillan, 2007

at the relatively narrow census block level.

Several studies have used similar approaches and the common range of calculated school quality premiums is a 2% to 5% increase in home prices resulting from a standard deviation increase in school test scores (Black, 1999; Bayer, Ferreira, McMillan, 2007; Kane, Staiger, Samms, 2003; Davidoff, Leigh, 2008). In all cases, researchers used school boundaries and compared entire neighborhoods on either side of the boundaries. The primary limitation of such studies is the mechanism controlling for housing characteristics. By relying on a linear model to control for differences in home characteristics across school boundaries, these studies are not robust against non-linear or non-additive valuations of home characteristics. So, for example, if adding extra bedrooms displays diminishing marginal return after the fourth bedroom, a linear model risks inaccurately valuing the extra bedrooms beyond four.

The central contribution of this study is a unique pairing methodology which more robustly controls for the problem of confounding home characteristics. Rather than compare entire neighborhoods, I pair individual home sales based on significant observable home traits such as bedrooms, square footage, and distance between potential pairs. Sale price and school quality are omitted from the pairing because they are the independent and dependent variables of interest. By only considering pairs which are highly similar across the pairing variables, and lie within 0.4 miles of each other, I control for differences in both observable home characteristics and unobserved neighborhood characteristics. This method does not rely on linear or additive valuation of home characteristics, so scenarios such as the example of diminishing returns for extra bedrooms will not impact my results.

Another key difference between this study and similar ones is the use of school districts, as opposed to individual schools, as the boundary region under consideration. In cases where district and city boundaries coincide, this approach has the potential to mistakenly attribute preferences for discretely changing city characteristics to the impact of school quality differences. However, my sample contains almost twice as many school districts as it does cities, indicating that boundaries often do not coincide. Therefore, this limitation has a minimal effect on my results. The benefit of using district boundaries is that I can consistently estimate the quality of *all* the public schools associated with a home. This difference in boundary specification also expands the scope of previously conducted research in the field.

I incorporate a unique methodology to a research question that has been well tested and find that my results are consistent with those of previous studies. My more robust control for differences in home characteristics across attendance boundaries contributes a higher level of confidence in the range of results already established in the field. In section II, I elaborate on the nature of the data collected and analyzed in this study. In section III, I specify the exact pairing process conducted. In section IV, I present my findings, and summarize them in context in section V.

## II. Data

To determine the school quality premium in real estate prices, data is necessary on both the quality of schools and the sales characteristics of the homes associated with those schools. My discussion of the data is presented according to these two categories.

### A. School Quality

The Academic Performance Index (API) is published by the California Department of Education twice a year. Scores range from a minimum of 200 to a maximum of 1,000; the general goal being to score above 800. The API score is released as both a Base score in spring and a Growth score in fall. Both annual releases are computed using the same criteria and scale, allowing interested parties to consistently measure the progress of a school or district over a single year. API calculation methodology does vary across years, however, so comparison across multiple years is discouraged[6]. While bi-annual API reports collect data on a wide range of school performance characteristics, the assigned API scores are based solely on test performance.

Scores on the Standardized Testing and Reporting Program (STAR) and the California High School Exit Examination (CAHSEE) tests are weighted individually for each school depending on grade level and attendance data. The STAR tests are conducted in grades two through eleven, and are the sole testing metric in grades two through eight. Topics tested include English Language Arts, Mathematics, Science, Social Science, and Life Science. CAHSEE tests are only taken in grades ten through twelve, and thus are only considered in high school API scores. Topics focus on English Language Arts and Mathematics.

[6] Education Data Partnership, (2013). Understanding the Academic Performance Index (API). Retrieved from http://www.ed-data.k12.ca.us/pages/understandingtheapi.aspx .

This study relies on the API as the sole metric of school quality[7] both because test scores have been shown to be the most significant consideration for parents (Hayes, Taylor, 1996) and because the API is the most widely distributed metric of school performance in California. Hayes and Taylor showed that, by expressing their preferences through real estate sorting, parents value test scores more than other school characteristics such as spending. Parents and economists agree that the best measure of school quality is test scores[8]. The API summarizes test results for schools and districts in California, and is the most easily accessible metric for prospective homebuyers. The API score can also be considered a proxy for a wider set of school quality characteristics that parents may value, making it an ideal metric for the purposes of this study.

I focus on district-wide API scores, rather than individual school scores, for both theoretical and practical reasons; though district wide scores do include some drawbacks which I will discuss. When parents evaluate the public school quality associated with a potential home, they are interested in all of the schools their child would attend as a resident of that home. A home within a highly rated elementary school zone, but a poorly rated middle or high school zone, has less educational value associated with it than the same home would have if the middle and high schools were also rated highly. District-wide scores allow me to use a single school quality metric for all homes in the same school district; meaning that all levels of K-12 education are considered in each home's API score. Furthermore, some school districts in our sample, such as Berkeley Unified, do not determine school assignment within the district solely on residential location.

[7] California Department of Education (2013). Data Quest, District API. Retrieved from http://api.cde.ca.gov/reports/page2.asp?subject=API&level=District&submit1=submit .
[8] Neighborhood School Characteristics: What Signals School Quality to Homebuyers? Hayes, Taylor, 1996

These differences pose problems with individual school assignment but do not influence district-wide data.

The drawback of this approach is that most school districts in my sample have several schools at each level of education. To the extent that the highly informed parent is only interested in the quality of the specific schools within the district their child will attend, applying the district-wide score to all homes within is an important generalization to note. However, because the district API is an average of the APIs of schools within the district, the best predictor of an individual school's API given the district API is the district API. While there may be differences in the quality of schools in the same district, district-wide scores are an unbiased estimate of the quality of schooling for all homes within the district.

In practice, the association of homes to school attendance areas is much simpler on the district level due to the higher accessibility of district boundary maps. Specific school boundary maps are often available only on an individual basis and in non-electronic formats. Collecting these boundaries for each school's attendance region and merging all of these maps into a single shapefile for use in GIS software was an impractical approach, given the limited resources of this study. District boundaries on the other hand are readily available on the census bureau website in shapefile format[9], simplifying the task of assigning individual homes to school districts.

The specific API scores used in this study are the 2003 API Base district scores (released in March of 2004) and the 2004 API Growth district scores (released in August of 2004). Each home sale observation in my dataset is assigned either the Base or Growth

[9] United States Census Bureau, (2013). UNSD (Unified School District), SCSD (Secondary School District). Retrieved from http://www2.census.gov/geo/tiger/TIGER2011/ .

score as its API value depending on whether the home was sold before or after the August Growth Report was released, respectively. This period of time was chosen, as explained below, to correspond to the most recent home sales data that avoided the impacts of the 2006 decline in the housing market.

### B. Home Sales

Real estate data was gathered through the extensive Multiple Listing Services (MLS) database made accessible by the Re/Max Accord real estate office, located in the East San Francisco Bay Area (East Bay). Data was available across all recorded home sale characteristics for all homes sold in Alameda and Contra Costa counties in 2004 and 2005. Unfortunately, data on rental properties was not available, and was thus omitted from this study. Furthermore, the inability to access reliable data from other counties in the San Francisco Bay Area served as the functional limitation to the geographic scope of this study.

Specifically, homes sold between April 1, 2004 and February 28, 2005 were exported from the MLS database. This is the period during which the 2003 API Base score and the 2004 API Growth score were the most recent metrics of school quality available to home buyers. I balanced threats against external validity by choosing the one-year period closest to the present while remaining uninfluenced by the potential distortions of the Great Recession. My goal was to produce a result that is as relevant as possible to both the present and to normal economic conditions.

The characteristics initially extracted from the MLS database—prior to any econometric analysis—included MLS number (a unique identifier), address, list price,

sale price, closing date, home square feet, lot square feet, bedrooms, bathrooms, partial bathrooms, garage spaces, year built, pool, stories, and building type (detached, condo etc.). Both the selection of relevant variables within this set and the deletion of questionably accurate entries are explained in the next section.

## III. Methodology

The design of this study was inspired both by the regression discontinuity framework of previous work in the field and by the nature of the data available.

Regression discontinuity design is a research method used to compare observations randomly placed on either side of a well-defined boundary. The randomness of the boundary variable allows researchers to treat the boundary as a natural experiment, separating control and experimental groups by the independent variable of interest. Selecting a smaller range around the discontinuity increases the internal validity of the natural experiment because the argument of randomization is more robust. A larger range around the discontinuity, on the other hand, results in a larger dataset and smaller prediction error.

Through the extensive pairing methodology explained in parts C and D of this section, I adapted the regression discontinuity design to a geographic boundary setting. Specifically, I accounted for the non-randomness of home placement along a boundary by strictly controlling for observed differences in home characteristics. Furthermore, those factors that were unobservable in home sales (primarily neighborhood characteristics) were controlled for by only comparing homes geographically near each

other; functionally selecting a narrow range around school district boundaries. Because my initial dataset contained over 33,000 home sales, I was able to limit the range around district boundaries while maintaining a significant sample size.

Ideally, the regression discontinuity design applied to this research problem would compare homes on opposite sides of a residential street, the middle of the street being the school district boundary. This would be ideal because the main unobservable characteristic in this design is the quality of neighborhood amenities. These amenities can include public facilities such as parks and swimming pools as well as services such as post offices and entertainment attractions. Access to these amenities generally changes smoothly across geographic borders. This allows for the assumption that with a sufficiently narrow band around a boundary, these unobserved amenities are evenly distributed among houses on either side. Applying this assumption I argue that, after controlling for differences in home characteristics, any difference in the price of homes immediately surrounding a district boundary can be attributed to the difference in school quality.

An overview analysis of the East Bay showed that there were almost no cases where a school district boundary is defined along a residential street. Therefore, the best way to ensure that homes being compared were in similar neighborhoods, while managing over 33,000 observations, was to impose a maximum distance requirement between potential pairs.

Given the nature of my home data as individual observations of home sales, and the functionality of the research discontinuity design, I decided to pursue a pairing process to identify highly similar homes across school district boundaries. This process

avoids the potential confounding effects of systematically different home characteristics across district boundaries. I first made all possible matches between homes within a broader band of district boundaries. Applying a similarity criterion and limiting geographic distance between homes to 0.4 miles, I narrowed my sample to contain only highly similar pairs of homes.

There were no school districts with identical school quality scores in my sample, so separation by district boundary guaranteed a perceivable difference in school quality. Finally, by analyzing the relationship between API score and the log of Sale Price within this paired subsample, I determined the portion of the difference in home sale prices that is caused by discrete differences in school quality. Through this process, I isolated the school quality premium in the housing market. Aggregating many of these matched homes, I was able to estimate an overall premium on education in the East Bay.

### A. Merging Datasets

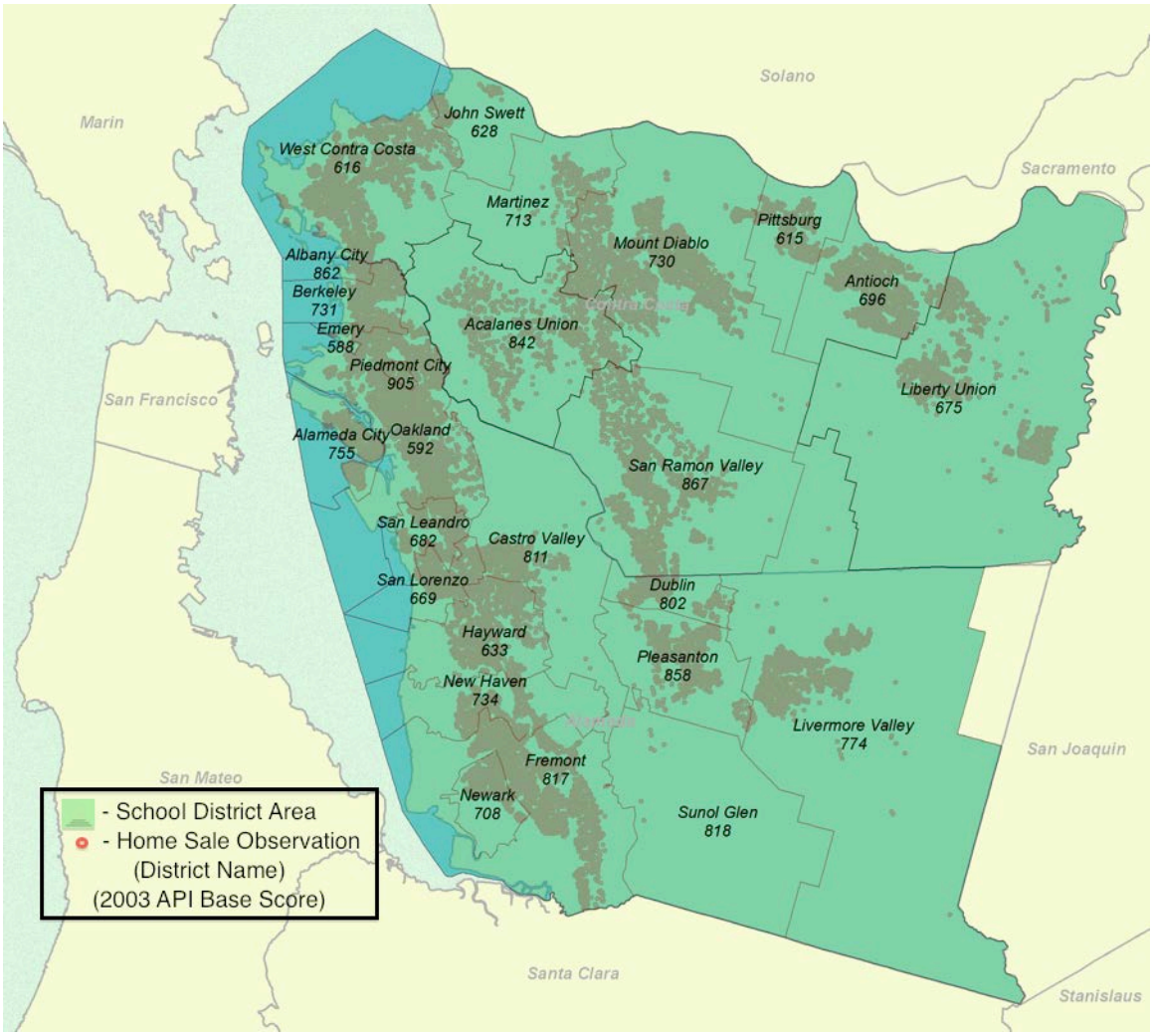
After collecting the relevant data, the first step was to combine all of the data into the home sale observations. The three datasets merged were the home sales data from the real estate database, the school district map files from the Census Bureau, and a database of API scores for school districts in the sample area.

I began by geocoding all housing observations according to street address using GIS mapping software. Over 95% of homes were successfully located and plotted in this geocoding process, incorporating latitude and longitude coordinates to the vast majority of home sale observations in my sample. I then added school district boundaries to my map using unified and secondary school district boundary shapefiles for California.

Finally, I merged a dataset with 2003 Base API scores for each district in the sample with the district boundaries layer, assigning a Base API score to each geographical school district region in my map.

Once this data had been successfully added to the map, I superimposed the layers such that each housing observation would include not only geographic coordinates, but also school district association and the resulting Base API score. After extracting this combined dataset from the GIS software, I replaced the 2003 Base API scores with 2004

**Figure 1: Home Sale Observations Mapped Against School District Boundaries & 2003 Base API Scores**



Growth API scores for all homes sold after the release of the Growth report. This produced a dataset suitable for econometric analysis.

### B. Data Cleaning

A small portion of the data collected from the MLS database contained extreme outliers in home characteristics that suggested data entry errors. In order to produce the most statistically relevant dataset it was necessary to trim any outliers.

Analyzing this problem through the lens of regression analysis, I sought to formulate a working model to predict sale prices. Using this model I was able to identify homes that displayed highly unpredictable statistics as those with the highest absolute value residuals. These home observations were the most likely to be caused by data entry mistakes, and were accordingly removed from the dataset.

To determine the relevant variables in this housing model, I first regressed log of Sale Price on all other observation characteristics and several plausible interactions and quadratic terms. The outputs of this regression can be seen as Model (1) of Table A. To determine which of these variables were relevant I kept only those displaying coefficients statistically different from 0. The reduced set of variables and the associated coefficients can be seen in Model (2) below. All prices were seasonally adjusted to correct for market variation over time.

Once the working model (Model 2) was specified, I calculated a predicted log Sale Price for each home. Using this prediction and the actual log Sale Price I constructed residuals, the difference between actual and predicted log Sale Price. These residuals allowed me to identify observations that exhibited prices highly different from what

**Table A: Developing a Working Model for Home Prices**

| | (1) log(Sale Price) | (2) log(Sale Price) | (3) log(Sale Price) |
|---|---|---|---|
| Home Square Feet | 0.0456*** | 0.0456*** | 0.200*** |
| | (0.00114) | (0.00114) | (0.00199) |
| Home Square Feet Squared | -0.0000457*** | -0.0000456*** | -0.000200*** |
| | (0.00000117) | (0.00000117) | (0.00000199) |
| Bedrooms | 0.207*** | 0.206*** | 0.216*** |
| | (0.00700) | (0.00676) | (0.00478) |
| Bedrooms Squared | -0.0154*** | -0.0155*** | -0.0230*** |
| | (0.000995) | (0.000974) | (0.000693) |
| Bedrooms x Stories | 0.0160*** | 0.0173*** | 0.00160** |
| | (0.00215) | (0.000820) | (0.000590) |
| Year Built | 0.00422*** | 0.00423*** | 0.00348*** |
| | (0.0000769) | (0.0000765) | (0.0000539) |
| Pool | 0.0675*** | 0.0676*** | 0.0336*** |
| | (0.00388) | (0.00388) | (0.00257) |
| Detached | 0.303*** | 0.303*** | 0.242*** |
| | (0.00489) | (0.00485) | (0.00316) |
| Condo | -0.123*** | -0.123*** | -0.118*** |
| | (0.00591) | (0.00591) | (0.00379) |
| Stories | 0.00428 | | |
| | (0.00665) | | |
| Closing Date | 0.000268 | | |
| | (0.000141) | | |
| Bathrooms | -3.58e-10 | | |
| | (7.91e-10) | | |
| Partial Bathrooms | -4.48e-10 | | |
| | (1.45e-09) | | |
| Garage Spaces | 1.09e-10 | | |
| | (5.22e-10) | | |
| Lot Square Feet | 1.94e-09 | | |
| | (1.59e-09) | | |
| Observations | 33023 | 33023 | 28642 |

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
In all models City is absorbed and Sale Prices are seasonally adjusted
Square feet terms are measured in units of 1,000

would be expected based on their observable characteristics. For example, this model predicted that a 1200 square foot, 2-bedroom, 2-bath, single story detached home in Castro Valley that transacted in April 2004 would have sold for approximately $550,000 on average. However, one observation in the dataset with these characteristics had a listed sale price of $4.25 million, more than 7.5 times the prediction. This observation was likely to have been miscoded. To eliminate such miscoded data, I deleted observations in the top and bottom 5% of the distribution of residual sale prices. I also deleted observations for which there were missing values. Model (3) in Table A shows the working model coefficients after removing entries suspected of being miscoded. All of the previously significant variables remained highly significant and the number of observations decreased by slightly over 10%. While coefficient values changed significantly after dropping 10% of the data, analysis of an alternative 15% threshold showed coefficients almost identical to those under the 10% specification, revealing insensitivity to the arbitrary threshold. This result shows that the shift in coefficients between Models (2) and (3) accurately reflects the removal of miscoded data.

### C. Determining The Best Pairs

The aim of this research design is to produce a reasonably large sample of the best possible pairs of home sales in order to examine the relationship between school quality and home prices among those pairs.

Ideally, I would consider homes that are identical on all dimensions except their school district locations, that sell in the same months, and that are extremely close to each other. However, this would yield too small of a sample. It was therefore necessary to

consider comparisons of houses that are similar but not completely identical. One way to derive such pairs would be to arbitrarily assign acceptable ranges of variation for each home characteristic, discarding potential pairs where the difference in any dimension exceeds the acceptable range. The drawback of this method is the arbitrary assignment of acceptable variation, and the inability to form tradeoffs between home characteristics and distance between potential pairs.

Allowing for tradeoffs between home characteristics and distance in my pairing process enabled me to both penalize a potential pair for being farther apart, and calculate the overall similarity of two homes, rather than setting arbitrary ranges for individual characteristics. The first result ensured that increasing the distance between homes would decrease their comparability. The second result increased the size of the dataset, thus increasing confidence in my final results.

To incorporate these benefits into my research design, I constructed a similarity formula using regression analysis. This formula determines the similarity of home sales across all characteristics, except educational quality. It uses coefficients on the difference in observable characteristics and on the distance between homes in a pair to define a limit of acceptability based on maximum composite distance.

The maximum geographical distance of the equation was 0.4 miles between homes, while the values of other home characteristics would function as tradeoffs with distance. Functionally, if a pair of houses was the full 0.4 miles apart, the differences in all other characteristics would have to be near zero for the pair to satisfy the similarity equation.

In constructing this formula (Formula 1) I first conducted a pairing process within individual districts, where district API scores are constant. To make this task computationally manageable, I randomly limited each district to 1,000 homes in expectation. I did this by creating a uniformly distributed random variable between 0 and 1 and only keeping observations where the random variable was less than 1,000 divided by the number of observations in the district[10]. With this limitation, the largest districts only created one million observations at the peak of the program's intensity, which was a task executable with household computing capabilities.

With these reduced district sets, I ran a pairing process within each district individually. All possible matches were made between houses that were within the same district and less than 0.7 miles apart. Given my goal of determining tradeoffs between distance and other home characteristics within 0.4 miles, including pairs more than 0.7 miles apart would not be useful.

The result of this pairing process was an individual dataset for each district that contained all possible pairs of homes within the district that were less than 0.7 miles apart. To determine the tradeoffs between home characteristics and distance, I generated an absolute difference variable for each variable in my dataset (all of the variables from Model (2) in Figure A, my Working Model) and regressed the difference in log Sale Price on all of the home characteristic differences and the distance between each pair. The coefficients of this regression indicated the relative weights distance and differences in home characteristics have in determining a difference in sale prices[11].

[10] This method both shrinks the individual district datasets and avoids sorting the data in each district iteration.

[11] Each difference and distance discussed is observed for every home pair observation and measures the corresponding value between the two homes in the pair.

Formula (1) shows the construction of the similarity equation. Ω is the parameter for maximum distance allowed, which I set to 0.4 miles. $\partial$ measures the distance in miles between homes in a pair. ß is a vector of coefficients on each of the differences in home characteristics (values of ß are listed in Table B). These values are equal to the coefficients on the respective variables in the differences regression scaled by the coefficient on distance. Finally, Δ is a vector of home characteristic difference variables between a pair of homes.

**Formula (1) :** $\Omega \geq \partial + (ß)(\Delta)$

This equation was particularly useful when comparing homes in *different* districts because it determines the similarity of a home pair in all characteristics except school district quality. Based on my application of the regression discontinuity design, I argue that the only difference in price between nearby homes that are highly similar across all observable characteristics is driven by the difference in school quality. This equation helped isolate the school quality premium in housing by measuring the similarity of pairs of homes across district boundaries. Specifically, the similarity equation was useful in weighing tradeoffs between distance and other differences in home characteristic variables, allowing us to set a maximum distance of 0.4 miles for very similar homes and impose a tighter geographic distance threshold for more dissimilar homes.

### D. Pairing Homes in Neighboring Districts

Having compiled the full dataset and formed a similarity equation to identify the most comparable homes, the final step in the construction of the analysis sample was to pair similar houses between neighboring districts. By forming all possible matches across district boundaries and dropping any pairs that did not satisfy the similarity formula, I reached a final dataset comprised of pairs of highly similar houses. In this dataset each pair of houses was geographically separated by a school district boundary.

The pairing algorithm I wrote to form pairs between districts looped through each district, forming pairs with all neighboring districts. All pairs of homes within 0.7 miles of each other but located in different districts were formed and saved into a temporary dataset. The ordering of homes within each pair was intentionally random.

I computed a similarity score for each pair of houses, using Formula (1) and the coefficients from Table B. The similarity score was scaled in miles – a pair of houses that is identical on all dimensions would receive a score equal to the distance between the houses, while differences in characteristics would produce a higher score (with the magnitude of the effect depending on the size of the difference and on the relevance of the particular characteristics to sale prices, as estimated in Table B.) The sample was then narrowed to home pairs with scores under 0.4 .

The analysis dataset thus only included pairs of very similar homes, according to the dual criteria of home characteristics and geographic distance. Within this sample, home pairs near the 0.4 mile threshold of distance were highly similar in home characteristics, while those further below the distance threshold were allowed slight dissimilarities in characteristics. In other words, those pairs of homes that did not meet

**Table B: Weights in Similarity Equation**

| Difference Variable | Weight | Difference Variable | Weight |
|---|---|---|---|
| Difference in Square Feet | 17.6993 | Difference in Year Built | 0.0058616 |
| Difference in Square Feet Squared | -1.347646 | Difference in Pool | 0.9459568 |
| Difference in Bedrooms | 7.6351 | Difference in Detached | 6.753416 |
| Difference in Bedrooms Squared | -0.8381265 | Difference in Condo | 5.186647 |
| Difference in Bedrooms x Stories | -0.222328 | | |

Square feet measured in units of 1,000

the standards of similarity imposed in Formula (1) were dropped from the sample, leaving only the most similar pairs of homes.

Table C shows the average absolute differences in each variable both before and after Formula (1) was applied to the full set of pairs within 0.7 miles. After dissimilar pairs were removed (Final Matched Pairs), average absolute differences dropped sharply in all variables. This decline in average absolute differences shows that Formula (1) successfully reduced the dataset of all pairs to a final set of highly similar pairs.

**Table C: Average Absolute Difference Between Home Pairs**

| | All Pairs Within 0.7 Miles | Final Matched Pairs |
|---|---|---|
| Square Feet | 0.584 | 0.031 |
| Square Feet Squared | 2.005 | 0.133 |
| Bedrooms | 0.910 | 0.156 |
| Bedrooms Squared | 5.579 | 1.495 |
| Bedrooms x Stories | 2.577 | 2.061 |
| Year Built | 16.720 | 8.288 |
| Pool | 0.209 | 0.005 |
| Detached | 0.353 | 0 |
| Condo | 0.255 | 0 |
| API* | 91.315 | 81.443 |
| Predicted Sale Price | $138,071 | $5,970 |
| Distance | 0.527 miles | 0.260 miles |
| Observations | 102397 | 212 |

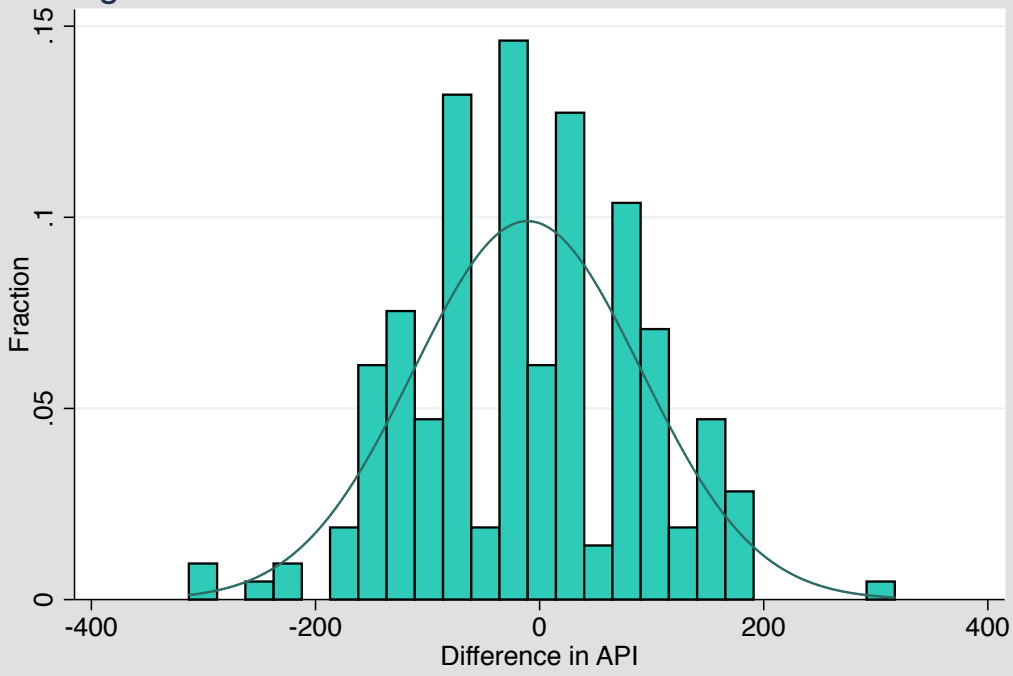
Square feet terms are measured in units of 1,000

*Statistics for non-absolute differences in API are displayed in Table D

## IV. Results

After finishing the pairing process, the dataset was composed of home pairs that were formed to be as similar as possible on all characteristics except school quality and log Sale Price. Table E shows that while some controls for differences in home characteristics were still significant in determining differences in log Sale Price (Model 1), the coefficient on school quality (API) did not vary significantly without these controls (Model 2). This shows that through the pairing methodology, I had limited the correlation between differences in home characteristics and differences in API.

Figure 2: Distribution of Differences in API Between Pairs




While there are always unobservable differences between homes, by limiting the distance between homes to 0.4 miles I only compared homes in the same or very similar

neighborhoods, overcoming any distortions due to neighborhood effects. Also, by maintaining over 200 pairs of homes in my final sample, I expect the differences across all unobserved home characteristics to be mean zero and have no significant bearing on my results.

**Table D: School Quality Difference Statistics After Pairing**

| | Mean | Median | Standard Deviation | Minimum | Maximum | Observations |
|---|---|---|---|---|---|---|
| Difference in API | -10.83019 | -15 | 101.547 | -313 | 317 | 212 |

The final analysis of the effect of API scores on home prices was implemented via a simple OLS regression of difference in log Sale Price on difference in API and difference in all home pair characteristics; this regression yields the final result of the study. I find that a 100 point difference in district API score, approximately one standard deviation (Table D), leads to a 3.3% difference in home prices (Model 1 in Table E); where the home in the 100 point higher scored school district is expected to be 3.3% more expensive than its pair in the lower scored district. Considering a major boundary in our sample, the 139-point difference between Berkley (731) and Oakland (592) API scores is predicted to have caused a 4.6% premium for Berkeley homes in 2004.

I also find that there is no extra value associated with "winning" the school quality contest against a paired home, which is not already captured by the valuation of API scores. This is shown by the insignificant coefficient on the variable "Higher School District Rating" in Model (1) of Table E. This implies that homebuyers are not concerned

**Table E: Final Results**

| | Similar Pairs Database | | | Naïve Model |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Difference in log(Sale Price) | Difference in log(Sale Price) | Difference in log(Sale Price) | log(Sale Price) |
| Difference in API | 0.000330* | 0.000369*** | 0.000449*** | |
| | (0.000128) | (0.0000776) | (0.0000864) | |
| Higher School District Rating | -0.013 | | | |
| | (0.0267) | | | |
| Difference in Square Feet | -0.955 | 0.259*** | | |
| | (0.589) | (0.0709) | | |
| Difference in Square Feet Squared | 0.157 | | | |
| | (0.121) | | | |
| Difference in Bedrooms | -0.288 | | | |
| | (0.195) | | | |
| Difference in Bedrooms Squared | 0.0336 | | | |
| | (0.0218) | | | |
| Difference in Bedrooms x Stories | -0.00112 | | | |
| | (0.00326) | | | |
| Difference in Year Built | 0.00290*** | 0.00293*** | | |
| | (0.000567) | (0.000531) | | |
| Difference in Pool | 0.0976 | | | |
| | (0.116) | | | |
| API | | | | 0.00189*** |
| | | | | (0.0000226) |
| Constant | -0.00219 | -0.00958 | -0.00626 | 11.71*** |
| | (0.0164) | (0.00775) | (0.0088) | (0.0166) |
| Observations | 212 | 212 | 212 | 28642 |

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Square feet measured in units of 1,000

with beating their closest neighbors in school quality, rather focusing on the absolute test scores within their district. Model (3) shows the results of a naïve regression of log Sale Prices on API in my original, unpaired database. The result of 16% clearly overstates the impact of school quality due to an abundance of confounding home and neighborhood characteristics.

To understand how the valuation of school quality differs across different socioeconomic groups, I divide my final sample into quintile groups by expected average log Sale Price of each pair. I repeated the specification in Model (1) of Table E for each of the five quintile groups; the results of which can be found ordered from lowest expected average log Sale Price to highest in Models (1) through (5) of Table F.

**Table F: Results by Expected Average Price Quintiles**

| | Quintile Groups By Expected Average log(Sale Price) | | | | |
|---|---|---|---|---|---|
| | (1) Difference in log(Sale Price) | (2) Difference in log(Sale Price) | (3) Difference in log(Sale Price) | (4) Difference in log(Sale Price) | (5) Difference in log(Sale Price) |
| Difference in API | 0.00161** | -0.000378 | 0.000208 | -0.000000379 | 0.000198 |
| | (0.000587) | (0.000301) | (0.000256) | (0.00027) | (0.000373) |
| Mean Sale Price | $411,354 | $443,676 | $488,854 | $557,702 | $725,756 |
| Observations | 42 | 43 | 42 | 43 | 42 |

Standard errors in parentheses

All models include the full set of variables used in Model 1 of Table E

$^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$

Interestingly, Model (1) of Table F shows that only the lowest quintile of the home price distribution is significantly sensitive to neighborhood school quality. This breakdown suggests the least expensive fifth of homes exhibit the highest price

sensitivity to changes in school quality. Using home price as a proxy for family income, I extrapolate that families with the least resources are the most willing to pay for public education through real estate sorting. Unfortunately, broken into quintiles my database became too small to confidently comment on preferences within the remaining quintile groups. Further experimentation with a larger dataset would be useful in gaining a better understanding of public education sorting trends by income levels.

## V. Summary

My hypothesis was that school quality is a significant positive component of real estate value, and that the real estate market serves as a socioeconomic barrier to high quality public education. I sought to prove this hypothesis by forming over 200 pairs of very similar homes across district boundaries and calculating the impact of education on sale prices within these pairs. I found that in the East San Francisco Bay Area—during the period when the 2003 API Base score and the 2004 API Growth score were the most recent measures of school quality—a 100 API point increase in district scores (approximately one standard deviation) led to a 3.3% increase in home prices.

Adjusting this estimate to the field of previous research studying elementary schools, I estimate a 3.76% increase in home prices resulting from one standard deviation increase in elementary school API scores[12]. This result is almost identical to the one produced by Bayer, Ferreira, McMillan, 2007 (3.75%)[13] and well within the 95%

[12] This difference in estimation is driven by the higher variation (SD=114, Mean=752) in elementary school scores I calculated using an alternate database of 2003 API Base scores for all elementary schools within my sample region. California Department of Education (2013). API Data Files, 2003 Base API – Data File. Retrieved from http://www.cde.ca.gov/ta/ac/ap/apidatafiles.asp .
[13] Comparison to Bayer, Ferreira, McMillan, 2007 is based on their estimation using only owner-occupied units, available on page 609 of their study.

confidence interval estimated by Black, 1999[14]. By finding results consistent with similar research while more strictly controlling for differences in home characteristics, this paper increases confidence that the estimated impact of school quality is not due to differences in home characteristics.

This study uniquely uses individual home sale observations and school district boundaries to develop a new and technically intensive pairing procedure to estimate the education premium in housing. My methodology more carefully considers home characteristics by formulating a similarity equation that weighs relative differences in home characteristics and removes pairs of homes that are not highly similar. By conducting my analysis on this final set of similar pairs of homes, my results are more robust to asymmetries in home characteristics across school boundaries, filling a void that had yet to be examined in the existing literature. This contribution, and the consistency of my estimates with the previous literature, reaffirms that public education in America carries a price.

I was able to use micro-data to extrapolate results about a major population center. While my scope is geographically limited to the East San Francisco Bay Area, I postulate that the education premium in housing calculated here is a good approximation for similar population centers across the United States.

This paper focuses on demonstrating that there is unequal access to high quality public education. I believe that the next step in addressing this problem lies in the growing field of experimental school reorganization and attendance policies. School districts around the country are serving as laboratories for innovative and hopeful methods of resolving the inequality of opportunity I, and other researchers, have

[14] I estimate the 95% confidence interval for Black 1999 to be [.18% , 4%].

identified. A constructive topic of future research would be to evaluate the relative merits of these various experimental policies. In order to formulate an effective national model for reform, it is first necessary to evaluate local efforts and adopt the best practices of successful school districts. As research, implementation, and policy efforts converge to a consensus in this field, I am optimistic that access to high quality public education in the United States will dramatically improve within the next 50 years.

## VI. References


- Mann, H., & Massachusetts. (1957). *The republic and the school: Horace Mann on the education of free men*. New York: Teachers College, Columbia University.
- Black, S. E. (May 01, 1999). Do Better Schools Matter? Parental Valuation of Elementary Education. *The Quarterly Journal of Economics, 114,* 2, 577-599.
- Bayer, P., Ferreira, F., & McMillan, R. (January 01, 2007). A Unified Framework For Measuring Preferences For Schools And Neighborhoods. *Working Paper Series,* 13236.)
- Kane, T. J., Staiger, D., & Samms, G. (September 15, 2003). School Accountability Ratings and Housing Values. *Brookings-wharton Papers on Urban Affairs, 2003,* 1, 83-139.
- Cropper, M. L., Deck, L., Kishor, N., & McConnell, K. E. (May 01, 1993). Valuing Product Attributes Using Single Market Data: A Comparison of Hedonic and Discrete Choice Approaches. *The Review of Economics and Statistics,* 2, 225.
- Tiebout, C. M. (January 01, 1956). A Pure Theory of Local Expenditures. *Journal of Political Economy, 64,* 5.)
- Bogart, W. T., & Cromwell, B. A. (June 01, 1997). How Much More Is A Good School District Worth?. *National Tax Journal, 50,* 2.)
- Lee, D. S., & Lemieux, T. (June 01, 2010). Regression Discontinuity designs in economics. *Journal of Economic Literature, 48,* 2, 281-355.
- Davidoff, Ian, & Leigh, Andrew (June 01, 2008). How Much do Public Schools Really Cost? Estimating the Relationship between House Prices and School Quality. *The Economic Record, 84,* 265, 193-206.
- Hayes, K. J., & Taylor, L. L. (January 01, 1996). Neighborhood School Characteristics: What Signals Quality To Homebuyers?. *Economic Review Federal Reserve Bank of Dallas,* 4, 2-9.

- United States Census Bureau, (2013). UNSD (Unified School District), SCSD (Secondary School District). Retrieved from http://www2.census.gov/geo/tiger/TIGER2011/ .
- Education Data Partnership, (2013). Understanding the Academic Performance Index (API). Retrieved from http://www.ed-data.k12.ca.us/pages/understandingtheapi.aspx.
- California Department of Education (2013). Data Quest, District API. Retrieved from http://api.cde.ca.gov/reports/page2.asp?subject=API&level=District&submit1=submit .
- California Department of Education (2013). API Data Files, 2003 Base API – Data File. Retrieved from http://www.cde.ca.gov/ta/ac/ap/apidatafiles.asp .