Abstract

The purpose of this project is to perform analysis on the price premium associated with properties that are situated nearby high quality state schools, which will be referred to as the good school premium (GSP). The analysis in this project is performed at the lowest possible level, to determine with maximum certainty whether the GSP exists. Every single individual house sale in England from 2017 to 2023 was included in the dataset, and for each sale, the postcode of that property was geo-located and compared to nearby schools in the area using a custom data analysis program that I developed specifically for this project. Both data on the closest primary and secondary schools, as well as weighted averages of all nearby schools (weighted by distance) were collected, and matched to each house sale. Highly significant results demonstrating the existence of the GSP for secondary schools, as well as its marginal decline over covid, were obtained. For primary schools, the results were borderline significant, which is likely due to a flaw in the instrumental variable used to estimate quality in primary schools.

In the process of examining the GSP, I will also present evidence that the 'urban exodus' due to covid is not statistically significant, bringing into question its validity.

Figure 1. Literature Review

Paper	Insight
1. Rosenthal (2003)[1]	Rosenthal considers a Hedonic price model, but only investigates secondary schools. She uses Ofsted scores a an IV for school performance (similar to my approach). I believe this choice of IV is flawed, since it doesn't remove the endogeneity sufficiently well. She also uses the closest school for each property, rather than a weighted average of schools, which may not adequately capture the effects of very good schools over a larger area. Like Rosenthal, I tried using Ofsted as an IV and got far more stable results for primary schools, but it had a first-stage $R^2 < 0.3$ and there is literature to suggest that it does not adequately capture school performance (vonn Stumm, 2021)[8]. Furthermore, Ofsted reports are from 1 to 4, and so do not capture academic quality with as much precision. This is especially important in the top end of the school quality distribution where we want to differentiate between the best schools.
2. Black (1999)[2]	Black uses a Hedonic price model to compare house prices on the border of school district areas in the U.S. This controls for a very significant amount of housing and social factors, but cannot be performed in England since England does not have strict school districts.
3. Cheshire and Sheppard (2004)[3]	They use a Hedonic model to determine GSP in Reading, England. They also find a stronger premium for secondary schools than primary schools, but a higher variance in the quality of primary schools. This supports my findings of less consistency between regions in for primary school attainment.
4. PWC (2019)[4]	PWC undertook a report on the premium associated with good schools, but does not seem to adequately control for any housing characteristics or reverse causality. Furthermore, they perform the analysis by postcode district rather than by distance, which also introduces inaccuracy.
5. UK Government (2017)[5]	The UK government investigated the existence of the GSP by postcode sector, similar to the PWC analysis. This does not take into account the actual geographical distance to the school, and nothing is used to control for reverse causality of more wealth in the area on academic performance.

A Model for House Prices

The issue with analysing the housing market on the macro level is the requirement to control for the significant reverse causality. Any aggregate measure (e.g. average wages, unemployment etc.) will be highly endogenous with house prices, since higher house prices cause the residents who move there to be wealthier, which correlates with more favourable socioeconomic characteristics. Furthermore, performing analysis on an inter region or inter local authority level would average out any local effects such as the GSP. Given the limited number of local authorities/regions and severe risk of endogeneity, this would make drawing any meaningful conclusions from such analysis impossible.

Most of the literature uses sub local authority level data to determine the GSP - and for good reason. While there are significant challenges obtaining good data at this level, it is absolutely necessary if significant results are to be obtained, and I concluded that the only way to get meaningful results was by performing an analysis on as low a level as possible.

When performing analysis at a per property level, we can more or less ignore macro effects (since they provide very little explanatory power for individual property price), and focus on the specific characteristics of each house. Since these are physical characteristics, the reverse causality inherent in socioeconomic variables is not present, as the price does not *cause* the physical characters of the house to change.

The Hedonic Price Model, used in this project, theorises that the price of some types of goods (or in our case, houses) are determined by their internal characteristics (e.g. number of rooms, distance to the city) and their external benefits (e.g. access to good schools). Essentially, these external benefits are derived from social goods, which do not have a direct price associated with them. Therefore, market equilibrium is achieved by movements in prices of that which grants access to these 'free' social goods. The Hedonic Price Model predicts that house prices should increase to reflect the additional utility to the homeowner of sending their children to a good school.

When using a Hedonic Price Model, the regression equation can be specified by splitting our explanatory variables into intrinsic characteristics of the property, $\mathbf{X_i}$, and the external characteristics - the educational quality of secondary and primary schools, Q^S and Q^P. The model is specified as:

$$P_i = \alpha + \boldsymbol{\beta}^\top \mathbf{X_i} + \gamma_S Q_i^S + \gamma_P Q_i^P + \epsilon_i$$

where P_i is the (deflated) sale price of property sale i. Note, there are no time subscripts - each i refers to a separate *sale*. This is because each house could be sold multiple times in a year, or not at all.

School Quality

Determining an appropriate way to define and measure 'quality' for the relevant schools is fundamental to this analysis. The direct measure, pupil attainment, is highly correlated with unobserved factors associated with a pupil's wealth and affluence, and is not just a function of the school's education quality. This means that when considering the effect of education on house prices, using attainment figures as a measure of quality would result in the problem of reverse causality, whereby higher house prices would attract a more wealthy population, leading to improved performance in the local schools due to the positive effect of wealth on attainment - a phenomenon described by Blanden (2004)[7].

Quality of school itself, Q_{it}, was measured by the total proportion of students who pass a certain academic success threshold. Due to the endogeneity described above, an IV is required. I used the proportion of students from disadvantaged backgrounds who succeed academically, *DisadvAttainment*, as the instrument for Q_{it} for both primary and secondary schools. This passes the relevance condition for its use as an IV, since an increase in academic achievement for the disadvantaged group will clearly increase the overall academic achievement achievement. Mathematically we can represent this as:

$$Q_{it} = \frac{a}{a+d} NonDisadvAttainment_{it} + \frac{d}{a+d} DisadvAttainment_{it}$$

where a is the number of non-disadvantaged pupils at the school, d is the number of disadvantaged pupils, *NonDisadvAttainment* is the proportion of non-disadvantaged pupils who succeed academically and *DisadvAttainment*, our IV, is the proportion of disadvantaged pupils who succeed academically. If *DisadvAttainment* increases, then it will directly increase Q, so the relevance condition holds.

The exclusion restriction also holds relatively well for *DisadvAttainment*, since there is no obvious way in which the academic success of disadvantaged pupils could impact house prices, except via the quality of the school (and thus the GSP). Unlike the overall academic performance, which suffers from reverse causality with respect to house prices, *DisadvAttainment* avoids much of this. An increase in more affluent pupils attending the school, and thus causing academic performance to increase, would be accounted for by *NonDisadvAttainment*. It would not be picked up by *DisadvAttainment*, as they would not qualify as disadvantaged pupils. There could potentially be a degree of reverse causality if the level of deprivation within the *DisadvAttainment* group is greater in areas with low house prices, since it may cause a lower *DisadvAttainment* score in those areas where there are many pupils that are deprived far beyond the threshold set for being considered disadvantaged. The other potential endogeneity issue is via the 'peer effect', where wealthy pupils have a positive impact on the performance of disadvantaged pupils through the positive academic and social attributes correlated with their wealth. This would be extremely challenging to control for and is likely small.

Methodology

I obtained data on all property sales across the whole of England by postcode from the UCL price by square meter data set[9]. I used computational methods (as described in the data table) to compare the geographical distance between every postcode of the property and the location of the nearby schools (also by postcode), and constructed several metrics for each record, which depend on the quality of local schools. In total, 5 092 169 data points were constructed, spanning 2017 to late 2023. School attainment data for a property was calculated using the attainment metrics from the school year starting a year before the property sale, so that the information was available to the buyer.

After extensive experimentation and consideration, measure settled upon was a weighted average of all the local schools (within 5km), weighted by distance from the property. This was chosen as it places postcodes on a continuum that differs per postcode and captures more variation in house price within areas with the same local schools. Using the closest school misses out on the variation, and does not take into account a slightly further away school that may still be easily accessible and is of higher quality.

The property sales were split into pre-covid, from 2017 to 2019 inclusive, and post-covid, from 2021 to 2023 inclusive. The year 2020 was left out due to distortions during the pandemic, and because the goal was to clearly identify the change from before to after. Since the data are in the form of individual house sales, it is technically not panel data, even though it is split over several years. As explained, every i represents a unique sale, and thus we cannot perform fixed effects or time series analysis on this data.

The following regression model was used to estimate the GSP, via the two stage least squares (2SLS) estimation method.

$$
\begin{aligned}
Price_i &= \gamma_0 + \gamma_1 Sec\hat{G}cse_i + \gamma_2 Prim\hat{R}wm_i + \boldsymbol{\beta}^\top \mathbf{X_i} + \epsilon_i \\
Sec\hat{G}cse_i &= \pi_0 + \pi_1 SecGcseDisadv_i + \pi_2 PrimRwmDisadv_i + \boldsymbol{\delta}_S^\top \mathbf{X_i} + u_i \\
Prim\hat{R}wm_i &= \phi_0 + \phi_1 SecGcseDisadv_i + \phi_2 PrimRwmDisadv_i + \boldsymbol{\delta}_P^\top \mathbf{X_i} + v_i
\end{aligned}
$$

2SLS regressions were run for all regions of England. Results for Greater London and North East regions (including Yorkshire and the Humber) were not included in the overall regression for England due to distortions in each region. The size and transport options in London remove some of the access barriers to schools in London, and create unique distortions. The North East had a significantly lower R^2 than the other regions, and it is not immediately clear why this is. which may be to do with unobserved interactions on the Scottish border, since Scottish schools are not considered. It may also be to do with the inclusion of Yorkshire and the Humber in the North East area.

Figure 2. Principle Data Sources	
1. UCL postcode level dataset[9]	Contains data on all property sales in the last two decades with the postcode included. Also includes characteristics of the house.
2. Department for Education School Performance Data (2017, 2018, 2019, 2022, 2023)	Performance data for KS2 and KS4 (seperate datasets) for all primary and secondary schools in England. This includes performance metrics, broken down by pupil demographic, including disadvantaged and non-disadvantaged.
3. Department for Education Ofsted Rankings (2017 - 2023)	Latest Ofsted rankings for every non-independent school in England. Ofsted rankings are not done every year, so each school's most up to date ranking is included.
4. England Index of Multiple Deprivation (IMD) 2019	Scores and rankings for all Lower Layer Super Output Areas in England, which measure different metrics of poverty and deprivation.

Data Processing and Geolocation Algorithm

Note on Data Processing: Due to the very large size of the data sets used and the time complexity of the operations performed (especially for the geolocation and comparison with schools/nearby towns), usual interpreted languages such as R or Python were not sufficient. Therefore I wrote a custom tool using the Rust programming language, which is a compiled, system level language, to efficiently perform the data processing required for this project. Several optimisation techniques (including memoization of geolocation data and multithreading) were used to speed up processing times (takes around 30 minutes to run, down from over 3 hours), since the time complexity is $O(nm)$ where n is total number of house sales and m is total number of schools.

Task	**Description**
Aggregating Data per School	For every relevant year, every school record in data source (2) was matched with the corresponding latest Ofsted figures (source 3) which range from around 2017 to 2023 (due to the disruption of the pandemic). Ofsted ratings are relatively consistent over time per school, especially with good performing schools, so this is not a significant issue.
Geolocating and Comparing Postcodes with Nearby Schools	Relevant property sales were determined (criteria: property located in England, sale occurred between 2017 and 2023) from source (1) and each property was geolocated using the geo_rust Rust library, which performs postcode lookups via the GeoNames[10] online geo-spatial directory. Distances from London, and from the nearest big city were calculated. For every school in the country, another lookup was performed on the postcode of the school, and distances were compared from the property to the school. The aggregated school data for the closest primary and secondary schools to the property postcode were associated with each property sale record. Additionally, a weighted average of attainment and Ofsted rankings for each school was also included, weighted by the relative distance from the property. Weights are w_s, where $w_s = \frac{5-d_s}{5}$, where d_s is the distance of the school from the property. For example, $SecGcse = (\sum_s w_s GcseScore_s)/\sum_s w_s$ for all s schools closer than 5km to the property, where $GcseScore_s$ is the proportion of pupils achieving 4-9 in Maths and English GCSEs in school s.
Matching postcodes with IMD	A lookup was performed using the ONS postcode to Lower level Super Output Area (LSOA) lookup[12] (since the IMD data is per LSOA), and then the IMD values for that LSOA were associated with each property sale record.

The full data set created for this project can be provided on request, as it is far too large to be included in this report.

Figure 3. Relevant Variables

Variable Name	Description	Details of Inclusion
Price	Price of the house for each sale, deflated by RPI inflation (base year 2017).	Dependent variable of the regression. Included as ln(*Price*) to account for the skewed distribution of house prices.
SecGcse	Weighted average of the proportion of pupils achieving grades 5-9 for Maths and English GCSEs in non-selective state schools closer than 5km to the property.	Used to as a measure of quality of secondary school, Q^S, since this directly captures overall academic performance. Endogenous with respect to *Price* due to reverse causality - higher house prices result in wealthier households in that area, which is correlated with better academic outcomes in local schools.
SecGcseDisadv	Weighted average of the proportion of disadvantaged pupils achieving grades 5-9 for Maths and English GCSEs.	Used as the IV for *SecGcse* since it is significantly less correlated with the part of *SecGcse* that is reverse-caused by house prices. Pupils are classified as disadvantaged and included in this measure if they qualified for free school meals in the past 6 years or have been in care/adopted.
PrimRwm	Weighted average of the proportion of all pupils making expected progress in reading writing and maths.	Used to measure quality of primary school, Q^P. Equivalent to *SecGcse* but for primary schools.
PrimRwmDisadv	Weighted average of the proportion of disadvantaged pupils making expected progress in reading writing and maths.	Used as an IV for *PrimRwm* quality of primary school. Equivalent to *SecGcseDisadv* but for primary schools.
FloorArea	The total floor space in m^2 of the property.	Included as ln(*FloorArea*) for elasticities and diminishing marginal utility to rooms.
NumRooms	Total number of rooms in the property.	Included as ln(*NumRooms*) for elasticities and diminishing marginal utility to rooms.
HouseAge	The approximate decade (specified as the first year of the decade) during which the house was built.	Included as ln(*HouseAge*) for elasticities.
DistLondon	Distance of the property from London.	Included as $\ln(DistLondon + 10)$. Constant added for stability near city centre. This is just used to control for any factors of the north south divide which could be endogenous with educational quality, so it is not important if the coefficient is difficult to precisely interpret. It just needs to capture the 'North South divide' which is well established in the literature.
SecDist	Distance of the property from the closest non-selective state secondary school.	
PrimDist	Distance of the property from the closest state primary school.	
IsFlat	Dummy for the property is a flat.	
IsTerraced	Dummy for the property is a terraced house.	
IsSemiDet	Dummy for the property is a semi-detatched house.	
EnvScore	The Index of Multiple Deprivation Living Environment score.	Included as ln(*EnvScore*) for elasticities. This measure is only determined by physical characteristics of the house (such as state of rooms, existence of central heating etc.), not endogenous socioeconomic attributes such as income.
DistCity	Distance of the property from the closest English city with a population of over 400 000.	Included as $\ln(DistClosestCity + 10)$ due to instability with properties near the city centre (as *DistClosestCity* approaches 0).

Figure 4. Results by Region

	Greater London		South East		South West		West Midlands		North West		North East		East Midlands		East of England	
Covid	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
SecGcse	1.959	1.202	0.743	0.862	1.069	0.897	1.011	0.832	1.073	1.040	1.431	1.262	1.016	1.007	1.029	0.739
	(0.012)	(0.010)	(0.006)	(0.007)	(0.013)	(0.012)	(0.013)	(0.014)	(0.012)	(0.011)	(0.018)	(0.022)	(0.013)	(0.014)	(0.012)	(0.011)
PrimRwm	-1.571	**0.076**†	**0.008**†	0.399	-0.248	0.245	**0.042**†	0.477	0.859	1.004	-0.704	-1.157	0.275	0.572	-0.494	0.074
	(0.025)	(0.024)	(0.012)	(0.013)	(0.017)	(0.016)	(0.022)	(0.022)	(0.021)	(0.022)	(0.024)	(0.038)	(0.022)	(0.024)	(0.017)	(0.016)
ln(*FloorArea*)	0.741	0.759	0.630	0.632	0.653	0.663	0.623	0.620	0.678	0.647	0.729	0.728	0.610	0.574	0.570	0.584
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.006)	(0.007)	(0.005)	(0.005)	(0.006)	(0.005)
ln(*NumRooms*)	0.050	0.067	0.070	0.092	0.053	0.069	0.087	0.104	0.145	0.169	0.148	0.178	0.132	0.157	0.119	0.127
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.006)	(0.007)	(0.005)	(0.005)	(0.005)	(0.005)
ln(*HouseAge*)	-2.893	-3.149	-1.018	-1.629	-0.625	-1.102	1.190	0.616	1.652	0.933	3.266	2.407	2.189	1.652	-0.888	-1.380
	(0.036)	(0.034)	(0.031)	(0.032)	(0.036)	(0.040)	(0.042)	(0.045)	(0.040)	(0.041)	(0.061)	(0.067)	(0.043)	(0.046)	(0.039)	(0.041)
ln(*DistLondon*)	-0.505	-0.469	-0.314	-0.236	-0.277	-0.189	-0.390	-0.348	-0.588	-0.587	0.128	0.022	-0.636	-0.548	0.044	-0.369
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.004)	(0.004)	(0.012)	(0.013)	(0.008)	(0.009)	(0.016)	(0.018)	(0.029)	(0.034)
ln(*SecDist*)	0.030	0.035	0.021	0.022	0.025	0.026	0.022	0.024	0.036	0.031	-0.019	-0.012	0.007	0.007	0.011	0.009
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
ln(*PrimDist*)	0.076	0.084	0.042	0.039	0.053	0.058	0.078	0.074	0.124	0.111	0.111	0.118	0.083	0.082	0.028	0.038
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.001)	(0.002)
IsFlat	-0.330	-0.405	-0.394	-0.482	-0.353	-0.433	-0.471	-0.585	-0.300	-0.442	-0.359	-0.475	-0.413	-0.538	-0.400	-0.475
	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.005)	(0.004)	(0.004)	(0.003)	(0.003)
IsSemiDet	-0.164	-0.172	-0.163	-0.176	-0.190	-0.202	-0.225	-0.222	-0.200	-0.206	-0.237	-0.265	-0.264	-0.267	-0.163	-0.170
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
IsTerraced	-0.224	-0.232	-0.261	-0.277	-0.280	-0.296	-0.388	-0.392	-0.483	-0.478	-0.491	-0.538	-0.448	-0.457	-0.275	-0.283
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
ln(*EnvScore*)	-0.052	-0.053	-0.026	-0.014	-0.022	-0.020	-0.047	-0.049	-0.057	-0.066	0.032	0.027	-0.004	-0.003	-0.023	-0.020
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
ln(*CityDist*)			0.030	0.062	-0.103	-0.083	-0.121	-0.112	-0.067	-0.088	-0.375	-0.409	0.010	-0.020	-0.571	-0.174
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.005)	(0.005)	(0.002)	(0.002)	(0.034)	(0.039)
Intercept	29.701	28.756	19.661	25.568	18.906	21.043	21.093	25.208	-5.000	2.823	-89.64	-86.88	17.640	24.814	13.882	16.100
	(0.385)	(0.368)	(0.256)	(0.259)	(0.295)	(0.330)	(0.380)	(0.400)	(0.362)	(0.372)	(1.578)	(1.718)	(0.535)	(0.577)	(0.334)	(0.354)
Observations	237945	236527	268311	247474	199739	174043	201104	168158	297679	259837	173746	154561	185236	157369	165059	152100
R^2	0.78	0.80	0.76	0.78	0.70	0.71	0.72	0.74	0.65	0.68	0.59	0.59	0.68	0.70	0.78	0.78

Values in brackets are the standard errors of the coefficient above.

† The p-value is **greater** than 0.001, thus the result is **not** significant. Given the large number of observations, we reject non-significance at a lower p-value. All other values are significant at the 0.001 significance level.

Note that Yorkshire and the Humber is included as part of the North East region.

Figure 5. Results for All of England Excluding Greater London and the North East

	2SLS Pre Covid		**OLS Pre Covid**		**2SLS Post Covid**		**OLS Post Covid**	
SecGcse	1.203	0.900	1.247	0.998	1.179	0.814	1.294	0.984
	(0.006)	(0.005)	(0.004)	(0.003)	(0.006)	(0.005)	(0.004)	(0.004)
PrimRwm	-0.733	0.047	0.931	0.772	-0.480	0.210	1.002	0.813
	(0.010)	(0.008)	(0.006)	(0.005)	(0.010)	(0.008)	(0.006)	(0.005)
ln(*FloorArea*)	0.752	0.676	0.723	0.664	0.749	0.662	0.722	0.653
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
ln(*NumRooms*)	0.012	0.070	0.015	0.071	0.045	0.094	0.044	0.093
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
IsFlat	-0.226	-0.325	-0.248	-0.333	-0.345	-0.424	-0.370	-0.433
	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
IsSemiDet	-0.221	-0.197	-0.224	-0.197	-0.224	-0.208	-0.226	-0.207
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
IsTerraced	-0.350	-0.345	-0.349	-0.342	-0.355	-0.358	-0.354	-0.354
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
ln(*EnvScore*)	-0.116	-0.031	-0.100	-0.021	-0.102	-0.031	-0.084	-0.021
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ln(*DistLondon*)		-0.612		-0.593		-0.554		-0.532
		(0.001)		(0.001)		(0.001)		(0.001)
ln(*CityDist*)		-0.020		-0.006		-0.032		-0.017
		(0.001)		(0.000)		(0.001)		(0.001)
ln(*HouseAge*)		0.644		0.726		-0.144		-0.055
		(0.018)		(0.017)		(0.018)		(0.018)
ln(*SecDist*)		0.024		0.020		0.024		0.022
		(0.000)		(0.000)		(0.000)		(0.000)
ln(*PrimDist*)		0.079		0.077		0.079		0.077
		(0.001)		(0.001)		(0.001)		(0.001)
Intercept	9.050	7.191	8.110	5.930	8.954	12.949	8.017	11.625
	(0.009)	(0.135)	(0.008)	(0.132)	(0.010)	(0.142)	(0.008)	(0.140)
Observations	1361263	1346733	1374373	1359699	1203552	1184805	1214451	1195542
R^2	0.49	0.70	0.53	0.70	0.54	0.71	0.57	0.71

Values in brackets are the standard errors of the coefficient above.

All values are significant at the 0.001 significance level.

Yorkshire and the Humber also not included in these data.

Figure 6. Robustness Testing

Note on large sample size: Due to the very large sample size involved, many of the standard test used are not very useful or meaningful. These tests are generally in the form of hypothesis tests under a certain null, H_0. The distribution under the null, Z_n, asymptotically converges in probability to some point θ_0 such that $Z_n \xrightarrow{p} \theta_0$. Therefore, by the definition of convergence in probability, for large n, the probability of observing the relevant sample estimate, $\tilde{\theta}_n$ under H_0 approaches zero: $\lim_{n\to\infty} \mathrm{P}(Z_n \geq \tilde{\theta}_n) = 0$, if H_0 does not *precisely* hold. Since it is almost impossible to find a perfect model, we will always reject the null for any reasonable significance level when n is large. Even so, it is still possible to use the actual test statistics to identify problems or improve the model, but we need to accept that the p-values will be negligible.

The following test results are from the fully specified 2SLS regressions for all of England excluding Greater London and the North East for both pre and post covid, but results are similar for all regressions.

Test	**Result**	**Analysis and Mitigation**
Heteroskedasticity	White test for Heteroskedasticity was run: $\chi^2(98) \approx 75000$, reject null of homoskedasticity.	Large statistic expected. Robust std. errors are used. Large sample size means minimal cost to standard errors of estimators using robust standard errors.
Normality of Errors	Due to large sample size, distribution of residuals was visually inspected and errors seemed to follow a normal distribution.	Distribution of errors plotted relative to the normal distribution and inspected visually, since it is difficult to interpret statistical tests with a high sample size. Estimators should still be asymptotically unbiased with non normal errors, even if OLS/2SLS is not the most efficient estimation method.
Collinearity	$1.18 \leq \text{VIF} \leq 3.56$, for all parameters. No significant multicollinearity.	Collinearity is manageable and should not cause significant issues. This test is still valuable for high sample sizes since it does not rely on a hypothesis test.
Model Misspecification	The Ramsay RESET test was performed with an F statistic of $F = 1400$ and a negligible p-value.	Once again, large sample size means that it is virtually impossible to pass the Ramsay RESET test at any reasonable significance level. However, the F statistic was reduced significantly from over 70000 by improving model specification and transforming some variables with logs.
IV: Weak Instruments	First stage R^2 is better for $Sec\hat{G}cse$ than $Prim\hat{R}wm$. $Sec\hat{G}cse$ we have $R^2 \approx 0.55$, $Prim\hat{R}wm$ we have $R^2 \approx 0.33$.	The low first-stage R^2 for $Prim\hat{R}wm$ suggest that *PrimRwmDisadv* may be a weak IV for primary school quality. Given that the results for primary schools are far less stable across regions, and less significant, this might help explain why. Further analysis in the analysis table.
IV: Endogeneity	Reject null hypothesis of exogenous variables. $\chi^2(2) \approx 18000$. P-values negligible.	Large sample size is an issue again here, but the extremely large χ^2 value still suggests that there is endogeneity without the use of the IV (provided that the IV correctly adjusts for this endogeneity). We can also observe the significant difference between the parameters when using 2SLS versus OLS, which suggests that the IV is indeed removing some endogeneity.

Figure 7. Analysis of Results

Question	Result
1. Does the data support the existence of the good school premium (GSP) for secondary schools in England?	Yes - the results for secondary schools in England are extremely significant and stable across all regions of England. The GSP can be observed most strongly in the Greater London region. However, this may be due to distortions that are unaccounted for. On average, for every 1% increase in the weighted average of percentage of pupils achieving grade 4-9 in Maths and English GCSEs in local secondary schools, there is a 0.90% increase in house prices pre covid, and a 0.81% increase post covid.
2. Does the data support the existence of the GSP for primary schools in England?	Borderline yes - there is significant variability of the coefficient on primary school attainment from the 2SLS estimator between regions and over time. This suggest that there are some distortions or unobserved effects that are interfering with the result. Further discussion below on why the IV for primary school attainment should be rejected. However the OLS coefficient is very significant and consistent across regions, and it would be surprising if all of this was due to reverse causality.
3. Does the data suggest that the GSP for secondary schools has changed over covid?	Yes - a slight decrease. We reject the null of no change at a 0.001 significance level. Decrease is observed in all but one region of England. I believe it is because the school quality gap has grown but house prices haven't responded (discussed in conclusion).
4. Does the data suggest that the GSP for primary schools has changed over covid?	Inconclusive - while we would reject the null of no change at a 0.001 significance level, issues with *PrimRwmDisadv* as an IV mean that it is more appropriate not to draw any conclusions.
5. Does the data support the existence of an 'urban exodus' since covid?	Not really. The coefficient on $\ln(DistLondon)$ increases in the Greater London region at a but is not significant. Some other regions experience an increase relative to their cities, via $\ln(CityDist)$, however the effect is small and in different directions between different regions. This result is in alignment with a report published by Atelier (2022)[13], which comes to a similar conclusion.

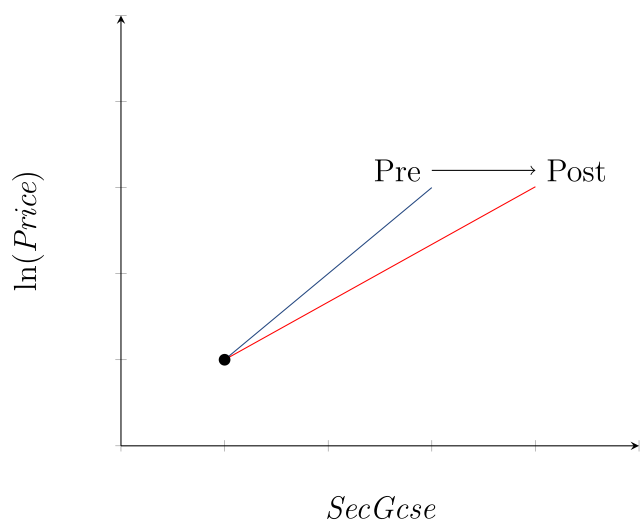
Analysis of Variables

Variable	Analysis
SecGcseDisadv as an IV	The results and tests suggest that using attainment figures for disadvantaged students as an IV for school quality was successful in secondary schools. As expected, we observe a higher school premium in the standard OLS regression, where we don't use the IV, since endogeneity is introduced via the reverse causality of high house prices on affluence and better academic attainment. One of the reasons for it's success is that the results are from official, external examinations, and so the data is reliable, objective, and comparable across regions. Furthermore, when the child sits GCSE's they are around 15/16, so are less influenced by parents/background, and more influenced by peers and the school. This means that the school quality quality is more able to have a positive effect on the pupil, so we can better observe the effects of the quality of education via this IV.
PrimRwmDisadv as an IV	The IV used to measure primary school attainment seemed to be weaker (low first stage R^2), and there seems to be little consistency between regions. This is likely due to 2 main issues: • Primary schools have significantly fewer pupils than secondary schools, and those pupils are more local. I found that many primary schools didn't actually report on disadvantaged pupils, likely because there weren't enough disadvantaged pupils at the school. This poses a serious selection bias problem if primary schools in more affluent areas are not included. This is likely to downwards bias the coefficient, since the more affluent areas are also more likely to have good primary schools (hence also the large difference between the 2SLS and OLS coefficients for primary schools). • There is much less transparency around how the Reading, Writing and Maths scores are determined at the primary school level. The tests used are marked in-house, and so there is always a question of validity of the figures, especially if there are incentives to publish certain results.

Evaluation and Conclusion

I will focus this conclusion on secondary schools, where my findings are most robust. It appears that the GSP has fallen slightly over the pandemic, which I suspect is a result of the *improvement* in school attainment (at least at GCSE level) post covid, only in the *best* schools. The mean weighted *SecGcse* score throughout England has increased from 0.613 to 0.636 from before to after the pandemic, and its estimated variance has also increased. Since most of this increase is an upward shift at the top end of the *SecGcse* distribution, I conclude that the fall in the observed GSP is due to an increase in the differences between the best performing and worst performing schools, while house prices did not react. In the figure below, each coefficient of interest is represented by the slope of its respective line. As we can see, when the gap between good schools and bad schools widens, the weighted *SecGcse* values for houses near good schools increases, while the actual price of the house remains fixed. Since these houses are more expensive on average (due to the GSP), the increase in GCSE scores will be experienced more for areas with higher house prices, thus lowering the slope of the post-covid line. Essentially, this means that each additional unit of quality causes a lower increase in house prices than pre-covid, because the educational quality of the best schools has increased relative to the worst schools, but house prices *have not changed* to reflect this.

Figure 8. School Quality Premium Pre and Post Covid



The only way to address the GSP is to improve educational standards in schools that are performing poorly. Perhaps incentive mechanisms can be introduced to reward high performing school for assisting poor performing schools to improve. Perhaps increased collaboration in the state education sector across local schools could help boost the schools which are underperforming, while not compromising on the high performing schools. Ultimately, it remains to be seen whether the premium will return to pre-pandemic levels, however, it is important that it is addressed in order to ensure equality of opportunity for all children, regardless of background.

References

[1] Rosenthal, L. (2003), The Value of Secondary School Quality†. Oxford Bulletin of Economics and Statistics, 65: 329-355. https://doi.org/10.1111/1468-0084.t01-1-00053

[2] Sandra E. Black, Do Better Schools Matter? Parental Valuation of Elementary Education, The Quarterly Journal of Economics, Volume 114, Issue 2, May 1999, Pages 577–599, https://doi.org/10.1162/003355399556070

[3] Cheshire, P. and Sheppard, S. (2004), Capitalising the Value of Free Schools: The Impact of Supply Characteristics and Uncertainty†. The Economic Journal, 114: F397-F424. https://doi.org/10.1111/j.1468-0297.2004.00252.x

[4] PWC, How does state school performance affect house prices in England? (2019). https://www.pwc.co.uk/economic-services/documents/how-state-school-performance-affects-house-prices-england.pdf

[5] UK Government, House prices and schools: do houses close to the bestperforming schools cost more? (2017), https://assets.publishing.service.gov.uk/media/5a82a832ed915d74e3402e69/House_prices_and_schools.p

[6] Mostafa, T. Decomposing inequalities in performance scores: the role of student background, peer effects and school characteristics. Int Rev Educ 56, 567–589 (2010). https://doi.org/10.1007/s11159-010-9184-6

[7] BLANDEN, JO, and PAUL GREGG. "FAMILY INCOME AND EDUCATIONAL ATTAINMENT: A REVIEW OF APPROACHES AND EVIDENCE FOR BRITAIN." Oxford Review of Economic Policy, vol. 20, no. 2, 2004, pp. 245–63. JSTOR, http://www.jstor.org/stable/23606627. Accessed 8 May 2024.

[8] von Stumm, S., Smith-Woolley, E., Cheesman, R., Pingault, J.-B., Asbury, K., Dale, P.S., Allen, R., Kovas, Y. and Plomin, R. (2021), School quality ratings are weak predictors of students' achievement and well-being. J Child Psychol Psychiatr, 62: 339-348. https://doi.org/10.1111/jcpp.13276

[9] UCL (2024), House Price per Square Metre in England and Wales, London Datastore, https://data.london.gov.uk/dataset/house-price-per-square-metre-in-england-and-wales

[10] GeoNames (2024), https://www.geonames.org

[11] MacKinnon, James G. "Using large samples in econometrics." Journal of Econometrics 235.2 (2023): 922-926.

[12] ONS, Postcode to Output Area Hierarchy with Classifications (May 2020) Multi-CSV Lookup in the UK, https://hub.arcgis.com/datasets/f923677199364b33a4212e0a475fee14/about

[13] Bob Pannell, The Covid city exodus: Reality, reversal or urban myth? (2022), Atelier, https://media.umbraco.io/atelier/mdfbanmt/atelier_tl_cities_whitepaper.pdf