Impact of Shadow Education on Educational Outcomes: Evidence from ASER 2016

Supervisor: Prof. Sonja Fagernas

Submitted in partial fulfilment of the requirements for the degree of MSc in Development Economics

By

Farhan Ajmaine

Department of Economics

University of Sussex

September 2024

Abstract

This study examines the impact of shadow education on educational outcomes in rural India using data from the Annual Status of Education Report (ASER) 2016. Employing a large sample of school-attending children aged 5-16, the research utilizes multiple regression techniques, including Ordinary Least Squares (OLS), family fixed effects, and ordered logit models, to analyse the relationship between private tutoring and academic performance in reading and mathematics. The methodology incorporates extensive controls for socioeconomic status, parental education, and regional variations through state-fixed effects while addressing potential endogeneity through family-fixed effects models. Key findings reveal a consistent positive association between private tutoring and improved learning outcomes across various model specifications. The study also uncovers substantial socioeconomic gradients in both tutoring receipt and educational outcomes, raising concerns about educational equity. Notably, gender effects are observed, with males generally performing better in mathematics and females in reading, while the impact of tutoring appears to be more pronounced for males in both subjects. Furthermore, the analysis reveals heterogeneous effects across subgroups, with tutoring having a more pronounced impact in government schools and at the primary level. These findings have important implications for educational policy and highlight the complex role of shadow education in shaping learning outcomes in developing countries, while also underscoring the persistence of gender disparities in educational achievement.

Table of Contents

Preface

I would like to express my gratitude to my supervisor, Prof. Sonja Fagernas, for her invaluable guidance, patience, and support throughout this research journey. Her insights and encouragement have been instrumental in shaping this work. I am profoundly thankful to my family for their unwavering support of my decision to pursue economics as a discipline out of passion. It is my sincere hope that this research contributes, even if in a small way, towards making education an engaging and empowering experience for all. May education kindle the flame of curiosity, transforming it from a mundane pursuit of credentials into a lifelong journey of learning and unlearning.

1.0 Introduction:

The beacon of knowledge casts a shadow, never the reverse. This shadow, born from the absence of light, stretches across the educational landscape. Yet, its reach, however far, cannot outshine its source. In the realm of formal education—our beacon—we find inspiration, questioning minds, and empathetic worldviews. But alongside it, a parallel shadow grows: supplementary learning that both complements and complicates. The complications arise when our view on education swings from learning to mere certificates to acquire. Credentials are a necessary first step to assess but without proper knowledge, their usage is in vain.

The motivation for this study stems from the growing recognition of shadow education's role in shaping educational landscapes and potentially exacerbating educational inequalities. This market for supplementary education represents a significant economic activity, with households often investing substantial portions of their income in pursuit of improved educational outcomes for their children. The theoretical underpinning of my research draws from the human capital theory, as pioneered by Gary Becker and others. This theory posits that education is an investment in human capital, yielding returns in the form of higher future earnings and improved life outcomes. The prevalence of shadow education can be seen as a manifestation of this investment behavior, with families seeking to augment their children's human capital beyond what is provided by formal schooling. However, this raises important questions about equity and efficiency in the education system. If significant educational gains are only achievable through private tutoring, what does this imply for students from less advantaged backgrounds who may not have access to such resources?

To investigate these issues, I have chosen to focus on India, a country with a large and diverse education system, where the shadow education phenomenon has gained significant traction. My primary data source is the Annual Status of Education Report (ASER) for 2016, a large-scale household survey that provides rich information on children's schooling status and basic learning levels in rural India. The ASER data is particularly valuable for this research as it employs a consistent methodology across years and covers a wide geographical area, allowing for robust analysis of educational outcomes and their determinants.

The core of my empirical strategy revolves around estimating education production functions, a standard approach in the economics of education literature. These functions model educational outcomes as a function of various inputs, including school characteristics, teacher quality, student attributes, and household factors. In my case, I'm particularly interested in isolating the effect of private tutoring on learning outcomes, as measured by reading and mathematics proficiency scores.

My analysis employs a series of econometric models to address potential sources of bias and to tease out the true effect of tutoring. I began with simple Ordinary Least Squares (OLS) regressions, progressively adding control variables to account for observable factors that might influence both the decision to seek tutoring and educational outcomes. These controls include demographic characteristics such as age and gender, socioeconomic indicators like parental education and household wealth, and geographical factors captured through state fixed effects.

However, OLS estimates may still suffer from endogeneity bias due to unobserved factors that influence both tutoring decisions and educational outcomes. To address this, I employed a family fixed effects approach, leveraging within-family variation in tutoring receipt among siblings. This strategy allowed me to control for unobserved family-level characteristics that might confound the relationship between tutoring and learning outcomes.

Furthermore, recognizing that the effect of tutoring might vary across different types of students and school settings, I estimated separate models for primary and secondary school students, as well as for government and private schools. This stratified analysis allowed me to examine potential heterogeneity in the returns to tutoring across different educational contexts.

In addition to these linear models, I also employed ordered logit models to account for the ordinal nature of the outcome variables, which represent different levels of reading and math proficiency. This approach allowed me to examine how tutoring and other factors affect the probability of a student achieving different proficiency levels, providing a more nuanced understanding of the impacts on educational outcomes.

A key feature of my analysis is the construction of wealth indices using principal component analysis (PCA) to capture household socioeconomic status. This approach, common in development economics research, allowed me to create a more comprehensive measure of household wealth than would be possible with simple income measures, which are often unreliable or missing in survey data from developing countries.

Preliminary results from my analysis reveal several intriguing patterns. First, there is a strong positive association between receiving private tutoring and improved learning outcomes in both reading and mathematics. This effect persists even after controlling for a wide range of individual, household, and geographical factors, suggesting that tutoring does indeed contribute to improved academic performance.

However, the magnitude of this effect varies across different subgroups. For instance, the impact of tutoring appears to be larger for secondary school students compared to primary school students, possibly reflecting the increased importance of academic performance as students progress through the education system. There are also interesting gender differentials, with boys seeming to benefit more from tutoring than girls in some specifications.

The socioeconomic gradient in educational outcomes is stark. Children from wealthier households and those with more educated parents consistently perform better on both reading and math assessments. This finding underscores the importance of family background in shaping educational outcomes and raises concerns about the perpetuation of inequalities through the education system.

Interestingly, the effect of tutoring remains significant even after controlling for these socioeconomic factors, suggesting that it provides benefits over and above what might be expected from a more advantaged family background alone. This could indicate that tutoring is serving as a mechanism for social mobility, allowing students to overcome initial disadvantages through additional educational investments.

The family fixed effects models provide particularly compelling evidence for the efficacy of tutoring. By comparing siblings within the same household who differ in their receipt of tutoring, these models control for unobserved family-level factors that might influence both tutoring decisions and educational outcomes. The persistent positive effect of tutoring in these models suggests that its benefits are not merely a reflection of unobserved family characteristics or selection effects.

From a policy perspective, these findings have important implications. On one hand, the positive impact of tutoring suggests that it could be a valuable tool for improving educational outcomes. This might argue for policies that expand access to tutoring services, particularly for disadvantaged students who might not otherwise be able to afford them. On the other hand, the very existence of a large shadow education sector could be seen as an indictment of the formal school system, indicating that it is failing to meet the educational needs of many students.

Moreover, the strong socioeconomic gradients in both tutoring receipt and educational outcomes raise serious equity concerns. If significant educational gains are increasingly tied to private investments in tutoring, this could exacerbate existing inequalities and limit social mobility. Policymakers may need to consider interventions that level the playing field, either by improving the quality of formal schooling or by providing targeted support to disadvantaged students.

As I delve deeper into this analysis, I aim to further unpack these complex relationships and their policy implications. By employing rigorous econometric techniques and drawing on rich household survey data, my research contributes to our understanding of the shadow education phenomenon and its impacts on educational outcomes in India. Ultimately, this work aims to inform policy discussions around how to create more equitable and effective educational systems in developing country contexts.

2.0 Literature Review:

Since the early 2000s, the “shadow education system” of private supplementary tutoring has expanded significantly worldwide. While this phenomenon has become global, it has traditionally been most prominent in East Asia. Japan’s juku and South Korea’s hagwons, which complement the regular school system for students of all ages, have long been well-known examples (Roesgaard 2006, Seth 2002). The United Kingdom uses the term “crammers” for similar post-school preparatory institutions. In recent years, this shadow sector has become increasingly visible not only throughout Asia but also in other regions around the world. This expansion reflects a growing trend in educational practices and result-driven assessment of educational outcome.

By definition, shadow education refers to a parallel educational system that operates alongside and supplements mainstream formal schooling. This phenomenon is primarily observed at the primary and secondary levels of education, with its most pronounced presence at the senior secondary level, followed by junior secondary and upper primary levels. While it also exists at post-secondary and pre-primary levels, its intensity and mechanics differ significantly at these stages.

In South Korea, it was predicted that 88% of elementary school students received tutoring in 2008. The percentage was 73% in middle school and 61% in regular high school (Kim, 2010, p. 302). In 2009, the average monthly cost of private education for each student was 242,000 won (equivalent to 242 US dollars) where 87.4% of students from elementary school, 74.3 percent from middle school, and 62.8 percent students from ordinary high school were using private tutoring services (Statistics Korea 2010).

Apart from East Asian countries, the prevalence of supplementary education can also be traced to South Asian regions. Nath's 2011 analysis of household survey data revealed that in 2008, private tutoring was widespread in Bangladesh. Among the receivers, 37.9% were primary students and 68.4% were secondary students. The practice was most prevalent among 10th graders, with over 80% receiving tutoring. It could be indicative of one of the first board exams students sit for after graduating high school. To aid the understanding, a separate survey in four Indian states found 58.8% of Grade 10 students received tutoring (Sujatha and Rani 2011). Indication of its 11cross11ion in primary education can also be seen in Sen (2010)'s paper where he showed 57% students were receiving private tutoring in the state of West Bengal in India.

The Pratichi Trust, founded by Amartya Sen, conducted surveys on primary education in West Bengal, India, in 2001/02 and 2008/09. While the follow-up study showed improvements in many areas, it also revealed an alarming increase in private tutoring dependence. The proportion of students receiving tutoring rose significantly, with 64% of standard primary school students and 58% of rural community school students now relying on it. Parental perception of tutoring as “unavoidable” increased to 78%, while economic constraints remained the primary barrier for non-participants. Sen criticized this trend, arguing that it exacerbates educational inequality, reduces teacher accountability, and undermines children’s right to quality elementary education. He noted that most tutoring content should have been covered in regular classes, highlighting systemic issues in the education system. This study provides crucial insights into the growing shadow education system in India, emphasizing its prevalence and potential negative impacts on educational equity and quality. Private tutoring availability varies significantly between rural and urban areas in India, as noted by Sujatha (2014). The exact causes of this disparity remain unclear, but likely stem from socioeconomic inequalities.

The effect of private tutoring is often times measured by student’s attention in class. Although attention in class is hard to measure, Lee’s (2013) study reveals a complex relationship between private tutoring and classroom engagement. While it shows a slight positive impact on students’ attentiveness, particularly among low-achievers who may gain confidence from extra coaching, there are also instances where students disengage from classroom activities, relying instead on their out-of-school support for exam preparation.

The ASER data suggests a 'divide' between students attending private schools or coaching classes and those who don’t, Wadhwa (2015) notes that only a small portion of this advantage is attributable to private education. Some researchers (Banerji and Wadhwa 2012; Desai et al. 2010) claim private inputs positively influence learning outcomes.

However, preliminary observations reveal that coaching centers often employ passive, instruction-driven pedagogy. Kumar (2012) argues that standard assessments fail to measure deeper learning, potentially masking curriculum and pedagogical shortcomings. Consequently, both schools and tutoring centers increasingly resemble test preparation facilities, potentially improving scores but discouraging imaginative learning. This raises critical questions about the true nature of academic achievement and the effectiveness of private tutoring in fostering substantive knowledge.

The popularity of supplementary tutoring stems from parents' belief that investing in education leads to better exam performance, access to prestigious schools, and ultimately higher lifetime earnings. In South Asia, the phenomenon of private tutoring has become a common occurrence showed by (Pallegedara 2011) study used national household survey data to analyse demand elasticity for private tutoring. The findings revealed a shift in perception: while private tutoring was considered a luxury in 1995/96, it had become viewed as a necessity by 2006/07.

Inadequacy in school quality also exacerbates the need for private tutoring. Teachers are often not accountable when they miss school however they put more effort into their tuition because there is a direct correlation between their income and effort (Chakraborty 2003). Moreover, the pedagogical shortcomings in schools are exposed when in a study 68 percent of the students said they attend tuition because they get to discuss exam papers and answers among peers. And 53 percent of them expressed the lack of school exercises compelled them to attend private tuition among other factors (Suraweera 2011, p.20–21)

According to several researchers, households—rather than instructors—are the ones driving the demand for private tutoring (Brehm and Silova 2014). A few studies suggest that parents occasionally ask instructors to offer extra tuition to students they currently instruct in school, especially in rural areas where there is still a dearth of activity in the tutoring industry. However, there are other times when educators subtly or overtly indicate to parents that their kids would benefit from coaching or, even more menacingly, that they wouldn't pass without individual instruction. Peer pressure and the demonstration effect are also present. Families of all classes find the promise of in-person tutoring, particularly in the area of English language instruction, to be quite appealing. Most of South Asia also exhibits similar viewpoints. According to Hamid et al. (2009:298), a student in Bangladesh made the following persuasive argument: "Private tutoring is needed because of the failure of schools in English teaching." There wouldn't be a demand for private instruction if English was taught well in schools.

The prevailing narrative that the tutoring sector is mostly driven by parental demand is challenged by the literature on private tutoring in India, namely in West Bengal and Tripura. Numerous studies indicate that this phenomenon is highly influenced by supply-side forces. Scholars have discovered proof that educators and tutors use calculated techniques to create demand for their services. The Shillong Times (2015) cites demonstrations in Tripura against a court order prohibiting private tuition as an example of this, where instructors are accused of pressuring parents and pupils to participate.

There is a general believe that shadow schooling improves academic performance because if it didn't, families wouldn't spend money on it. That presumption might not always be accurate, though. A lot relies on the learners' aptitude and drive as well as the calibre of the instruction. While some tutors are quite skilled, they may deal with pupils who lack motivation or academic ability. Consequently, some students are capable and driven, but their tutors are inexperienced with the subject matter and pedagogy. Alternatively, because the majority of their peers appear to be doing the same, children might keep going to tutoring sessions. Going back to Nath's household data in Bangladesh, indicate that 49.6% of pupils aged 11–12 who had received private tutoring met the benchmark criteria of having a basic education, while only 27.5% of students without tutoring met the benchmark drawing from a 1998 national survey data. Moreover, Hamid et al. (2009:293) reported on a survey of 228 grade 10 students in eight rural schools and found that students who had received private lessons had double the chance of attaining higher grades than their counterparts who had not received private lessons. However, both of these studies can be attributed to correlations rather than causation.

The research on how private tutoring affects mathematical achievement has produced contradictory results and methodological issues. Because of its thorough methodology, Kuan's (2011) study of 10,013 grade 9 pupils in Taipei, China, is very notable. Kuan discovered that kids receiving tutoring were typically more studious, higher achievers, and from higher social classes after adjusting for socioeconomic position, aptitude, and attitude. The study found that while achievement increased little, motivated students benefited more. Kuan's study was limited, though, in that it only examined one semester and combined all tutoring styles into a single variable, making it unable to analyse long-term effects or implications across all grade levels (Kuan, 2011, p. 353, 362).

Byun (2011) carried out a comparable study in the Republic of Korea. Byun compared the impact of tutoring on mathematical academic ability for a nationally representative sample of lower secondary pupils using propensity score matching. He discovered that there was a slight variation in the achievement improvements caused by cram education, which mostly focused on test preparation. Other tutoring methods, like one-on-one, online, and correspondence tutoring, had minimal impact. This somewhat confirmed the results of Kang (2009), who likewise discovered modest but beneficial benefits from tutoring investments based on 1,752 students' experiences monitored over the course of a Korean Education and Employment Panel longitudinal study.

In the Indian region, a study conducted by Aslam and Atherton (2011) examined information from the Uttar Pradesh and Bihar SchoolTells surveys conducted in 2007–2008. Four thousand pupils in grades two and four in 160 rural primary schools were the subject of the survey. Children who got tutoring improved in reading and math, with improvements being larger in public schools than in private ones. They also examined data from the Annual Status of Education Report (ASER- Pakistan 2011), which polled primary school-age children in 19,006 rural homes. Private tutoring has been found to benefit children from all socioeconomic backgrounds. Results were less significant in mathematics but particularly noticeable in reading scores.

According to Lee (2013), while private tutoring helps close the success gap in high school, it actually makes middle school educational inequality between "high" and "low" performers worse. In other words, high performers do not gain as much from private tutoring in high school as low achievers do. However, it is difficult to avoid suggesting that the privately paid supplement, determined by the size of the purse, creates new inequality along class lines and compounds the advantages of the upper middle classes given the widely apparent quality gap between various tutoring services and their instructional resources and study materials (Majumdar 2014).

According to (Dang and Rogers 2008), household spending on private tutoring has surpassed public sector education expenditures in nations like Turkey and the Republic of Korea. The perceived worthlessness of some of the content that tutoring students learn in order to prepare for university admission exams has been brought up in criticism of certain parts of the generally excellent educational systems in Korea and Japan—both of which have sizable private tutoring industries. Korea has undergone reform as a result of worries about the significant financial burden that tutoring places on parents (Kim 2001). The ratio of the cost of private tutoring per child to the per capita household expenditure provides a more accurate measure of the burden that private tutoring places on households. The average cost of private tutoring is 3.1% of household consumption expenditure per capita; however, if we limit our analysis to pupils who really paid for private coaching, the cost of private tutoring rises to 16.5% of consumption expenditure per capita (Azam 2006).

3.0 Data:

I came to know about The Annual Status of Education Report (ASER), a large-scale household survey conducted annually in rural India to assess children's schooling status and basic learning levels while reading Poor Economics by Abhijit V. Banerjee. The survey has two primary objectives: to obtain reliable estimates of children's schooling and basic reading and math abilities and to measure changes in these statistics over time. ASER employs a consistent set of core questions each year to track these fundamental metrics, while also incorporating additional questions to explore various dimensions of elementary and secondary education.

ASER typically utilises a two-stage sampling design to ensure representative data collection across rural India. In the first stage, villages are randomly selected from the Census village directory for each rural district. The 2016 ASER survey marked a transition to using the Census 2011 sampling frame, which became publicly available in 2015. This update ensured that the survey's sampling methodology reflects the most current demographic information available. In the second stage, households within the selected villages are randomly chosen for participation in the survey.

The motivation strategy is designed to generate a representative picture of each district, with all rural districts included in the survey. These district-level estimates are then aggregated to produce state and national-level data. The sample size for ASER is determined by several factors, including the incidence of the measured attributes in the population, the desired confidence level of estimates (set at 95%), and the required precision on either side of the true value.

Based on these considerations and the need for robust district-level data, ASER 2016 maintained a sample size of 600 households per district. This sample is divided among 30 villages in each district, with 20 households surveyed in each selected village. The selection of villages employed the probability proportional to size (PPS) sampling method, which ensures that villages with larger populations have a proportionally higher chance of being included in the sample. This approach helps to balance representation and maintain consistent probabilities of household selection across villages of varying sizes.

Within the selected villages, ASER employed a structured random selection process for households. Field investigators divide each village into four parts and select five households from each part using a systematic sampling approach. This method aimed to preserve randomness while ensuring coverage of different areas within the village, including peripheral households that might be missed by centralised selection methods.

ASER collected information on all children aged 3-16 years in the selected households and administers learning assessments to children aged 5-16 years. This household-based approach offers several advantages over school-based testing, including the ability to assess out-of-school children and avoid potential biases associated with school-based sampling. I am also using the household data to support my thesis.

To facilitate the measurement of changes in learning outcomes over time, ASER has adopted a rotating panel design for village selection. Typically, each year 10 villages from three years prior are dropped, 20 villages from the previous two years are retained, and 10 new villages are added. However, for ASER 2016, a fresh sample of 30 villages was drawn for each district due to the transition to the Census 2011 sampling frame. This refresh ensured that the survey's sample remains current and representative of the changing rural landscape.

3.1 Data Preparation:

In the data preparation phase of my research, I undertook several crucial steps to transform and enhance the raw data for analysis. My focus was on creating a set of variables that would allow me to thoroughly examine the relationships between socioeconomic factors and educational outcomes.

To begin, I generated a series of dummy variables to represent key binary characteristics in my dataset. These included indicators for tuition status, parental education levels, household amenities, and child gender. For each of these variables, I used conditional statements to create binary indicators and subsequently cleaned the data by replacing any missing values with null entries. This process ensured the integrity of my data while preparing it for more complex analyses.

A central component of my data preparation was the construction of wealth indices to capture household economic status. I approached this in two ways. First, I created a simple additive wealth index by summing binary indicators for several household assets and amenities, including four-wheelers, two-wheelers, mobile phones, newspapers, electricity connections, and reading materials. This provided me with a straightforward measure of household wealth on a scale from 0 to 6. To complement this, I also employed Principal Component Analysis (PCA) on the same set of variables to generate two additional wealth indices. This dual approach to measuring wealth allowed me to test the robustness of my findings to different specifications of household economic status.

To only keep the students who are attending school, I created variables to identify children currently in school and those not in school. The 'in_school' variable captures attendance at government, private, madarsa, or other types of schools, while the 'not_in_school' variable identifies children who have never enrolled or have dropped out. Using these, I defined my final sample selection variable, 'school_attendance_sample', which includes only those children currently attending school and not simultaneously categorized as out of school. The final number observations come to 443,468 from 639,752.

In my analysis, I generated dummy variables to distinguish between primary and secondary school students. This categorisation was crucial for understanding how the effects of various factors might differ across educational levels. I created ‘in_primary_school’ for students in classes 1 to 5, and ‘in_secondary_school’ for those in classes 6 to 12. This stratification allowed me to conduct more nuanced analyses, recognising that determinants of educational outcomes may vary significantly between primary and secondary levels of education. Recognising the importance of distinguishing between various educational institutions, I generated a set of dummy variables to represent government schools, private schools, madarsas, and other educational establishments. This approach allowed for a more granular examination of the characteristics and outcomes associated with each school category.

The study utilises two key outcome variables to assess children's academic proficiency: reading level (read_code) and arithmetic level (math_code). Both variables are coded on a 5-point ordinal scale, providing a nuanced measure of children's abilities in these fundamental academic skills.

The reading level variable (read_code) captures a child's reading proficiency across five distinct levels. A score of 1 indicates that the child could not read anything, representing the lowest level of reading ability. A score of 2 is assigned when a child can identify letters, showing the beginning stages of reading development. Children who can read words are given a score of 3, indicating an intermediate level of reading skill. A score of 4 represents the ability to read a Standard 1 level text, while the highest score of 5 is given to children who can read a Standard 2 level text, demonstrating advanced reading capabilities for their age group.

4.0 Methodology:

In my methodological approach, I extended my analysis to incorporate geographical variations by generating dummy variables for different states. This decision was motivated by the recognition that educational outcomes and characteristics often exhibit significant regional disparities. By creating these state-specific indicators, I aimed to capture and quantify the potential influence of state-level factors on the variables of interest in my study. I focused on eleven key states that were particularly relevant to my research questions and hypotheses. These states were Kerala, Tamil Nadu, Bihar, Uttar Pradesh, Maharashtra, Rajasthan, Madhya Pradesh, Gujarat, West Bengal, Himachal Pradesh, and Assam. I created a new variable to categorize these states, assigning each a unique numerical identifier. This allowed me to isolate these specific states for more detailed analysis while grouping the remaining states into a separate category. To facilitate more nuanced analyses, I generated dummy variables for each of these key states. This approach enabled me to examine state-specific effects and conduct comparative analyses across these diverse regions.

In my initial regression, I employed Ordinary Least Squares (OLS) regression methods to examine the relationships between various factors and student performance in reading and mathematics. This approach allowed me to quantify the effects of key variables while controlling for a range of demographic and socioeconomic factors.

For each model, I included the tuition dummy variable as the primary predictor of interest, alongside a comprehensive set of control variables. These controls helped account for various demographic and socioeconomic factors that might influence academic performance. The motivation behind including these variables in my regression models stems from the theoretical framework of the education production function and the complex interplay of factors that influence educational outcomes. This approach recognizes that learning is a multifaceted process, influenced by a variety of inputs at the individual, household, and school levels.

By incorporating child_age, I aim to capture the cumulative nature of learning over time, acknowledging that older students may have had more opportunities to develop their reading skills. The inclusion of the tuition_dummy variable represents a key schooling input, allowing me to examine how additional educational resources or interventions might affect reading proficiency. The male variable enables me to investigate potential gender differences in learning outcomes, which is crucial for understanding and addressing any disparities in educational achievement.

Household characteristics such as hh_type, hh_electricity, and total_member are included to represent the broader socioeconomic context in which learning takes place. These variables capture aspects of the home environment that may influence a child's ability to study and learn effectively. Furthermore, variables like hh_reading, father_edu, and mother_edu represent household investments in education and parental human capital. These factors are critical as they reflect both the direct support for learning at home and the potential for intergenerational transmission of human capital. Parents with higher education levels may be better equipped to assist their children with schoolwork or may place a higher value on education, potentially leading to better learning outcomes.

In my analysis, I first created a wealth index to capture the socioeconomic status of households. This index was constructed by summing several binary indicators of household assets and amenities:

wealth_index = hh_4wheeler + hh_2wheeler + hh_has_mobile + hh_has_newspaper + hh_electricity + hh_reading (1)

To refine this measure and account for potential correlations among these indicators, I then employed principal component analysis (PCA). This technique allowed me to create more nuanced wealth indices (wealth_index_pca1 and wealth_index_pca2) that capture the underlying variability in household wealth more comprehensively. The inclusion of wealth indices (wealth_index_pca1 and wealth_index_pca2) is motivated by the need to capture the overall economic status of the household, which can influence both direct investments in education and the overall learning environment. Wealthier households may be able to provide more educational resources, create a more conducive learning environment, or invest in additional tutoring or educational materials. By using principal component analysis to create these indices, I aim to capture a more comprehensive measure of household wealth that goes beyond simple income measures, which in the ASER dataset was missing.

In my research, the theoretical framework of my models is inspired by Paul Glewwe and Karthik Muralidharan (2015) paper's education production function framework, which provides a structural relationship between various inputs and learning outcomes. My approach can be represented by the following general equation:

$$A = f(S, Q, C, H, I) \qquad [2]$$

Where A represents academic achievement (in this case, reading proficiency and mathematical abilities), S represents schooling inputs, Q represents school and teacher characteristics, C represents child characteristics, H represents household characteristics, and I represents household investments in education.

Building on these foundations, I developed a series of regression models to examine the relationship between tuition and academic performance in reading and mathematics. The general form of these models can be represented by the following equations:

$$\mathit{read\_code\ (or\ math\_code)} = \beta_0 + \beta_1 \mathit{tuition\_dummy} + \beta_2 \mathit{child\_age} + \beta_3 \mathit{male} + \beta_4 \mathit{total\_member} + \beta_5 \mathit{wealth\_index\_pca1} + \beta_6 \mathit{wealth\_index\_pca2} + \beta_7 \mathit{father\_edu} + \beta_8 \mathit{mother\_edu} + \beta_9 \mathit{house\_type\_dummy} + \Sigma_i\, \beta_i\, \mathit{key\_state}_i + \varepsilon \qquad [3]$$

Equation [3] corresponds to Model 1 in my analysis. Here, read_code represents the measure of reading proficiency (A in the production function), child_age represents a key child characteristic I, tuition_dummy and male represent schooling inputs and child characteristics (S and C), while the remaining variables represent various household characteristics and investments (H and I).

In my analysis, I introduced interaction terms to explore how the effect of tuition on academic performance might vary across males and females. Specifically, I incorporated an interaction between the tuition dummy variable and the indicator for being male to capture potential heterogeneity in the effects of tuition across genders. This is crucial for understanding how educational interventions might need to be tailored to address potential gender disparities in academic performance.

This approach can be represented in the regression equation as follows:

$$read\_code\ (or\ math\_code) = \beta_0 + \beta_1 tuition\_dummy + \beta_2 (tuition\_dummy \times male) + \beta_x X + \varepsilon \qquad [4]$$

Where X represents the vector of control variables, and β_x their respective coefficients.

In equation [4], β_1 represents the effect of tuition for female students (when male = 0), while $\beta_1 + \beta_2$ represents the effect for male students (when male = 1). The coefficient β_2 itself indicates how much the effect of tuition differs for male students compared to female students.

I structured my analysis in several stages, progressively building more complex models. Initially, I ran basic models examining the relationship between tuition and academic performance, controlling only for the child's age. I then expanded these models to include a comprehensive set of control variables and conducted separate analyses for reading and mathematics scores, enabling me to identify any subject-specific patterns or differences in the effects of tuition and other variables.

To ensure a thorough understanding of how these relationships might vary across different educational contexts, I also performed separate analyses for different school types (government and private) and educational levels (primary and secondary). This stratified approach provided insights into how the effectiveness of tuition might differ based on the institutional setting and stage of education.

In a different model, I employed ordered logit models to further investigate the relationship between tuition and academic performance in reading and mathematics. This approach was particularly appropriate given the ordinal nature of the outcome variables, which represented different levels of reading and math proficiency. The equation for this model can be expressed as:

P(Y_i ≤ j) = logit^(-1)(α_j - β_1X_1i - β_2X_2i - ... - β_kX_ki) [5]

In the Equation [5], Y_i represents the ordinal outcome variable (reading code proficiency or math) for individual i, j denotes the different levels of the outcome, α_j is the intercept for each level j, and β_k represents the coefficient for each independent variable X_k.

To complement this analysis and provide more interpretable results, I also conducted marginal effects analysis (dy/dx) on these ordered logit models. This marginal effects analysis provided a more intuitive interpretation of the results. For instance, it allowed me to quantify how receiving tuition changes the probability of a student being in the highest reading or math proficiency level, holding other factors constant. These marginal effects are particularly useful for policy implications, as they provide a clear measure of the potential impact of interventions like expanding access to tutoring services.

I also implemented a series of household fixed effects models to examine the relationship between tuition and academic performance in reading and mathematics. This approach allowed me to control for unobserved household-level characteristics that might influence both the decision to enrol a child in tuition and their academic outcomes.

I began by setting the household ID as the panel variable, which enabled the fixed effects estimation. For both reading and math outcomes, I estimated three progressively complex models: i) a basic model with only tuition as the predictor. ii) an expanded model including time-varying controls such as child's age and gender. iii) a comprehensive model that included interactions between tuition and school level (primary and secondary).

$$READ/MATH_{it} = \alpha_i + \beta_1 TUITION_{it} + \varepsilon_{it} \qquad [6(i)]$$

$$READ/MATH_{it} = \alpha_i + \beta_1 TUITION_{it} + \beta_2 AGE_{it} + \beta_3 MALE_i + \varepsilon_{it} \quad [6(ii)]$$

$$READ/MATH_{it}= \alpha_i + \beta_1 TUITION_{it} + \beta_2 MALE_i + \beta_3 (TUITION_{it} \times MALE_i) + \beta_4 AGE_{it} + \varepsilon_{it} \qquad [6(iii)]$$

In Equation [6(i-iii)] READ/MATH_it represents the reading code proficiency for individual i at time t, α_i denotes the household fixed effect, TUITION_it is a dummy variable for tuition status, AGE_it is the age of the child, MALE_i is a dummy variable for male gender, and ε_it is the error term. The β coefficients represent the effects of the various predictors on the reading code proficiency. By using household fixed effects, I effectively controlled for all time-invariant household characteristics, such as parental education, socioeconomic status, and other factors that might be constant across siblings within the same household. This approach helped to isolate the effect of tuition on academic performance by comparing siblings within the same household who may have different tuition experiences.

In addition to examining the binary effect of receiving tuition, I also investigated the impact of the amount spent on tuition. In my analysis, I employed a series of ordinary least squares (OLS) regression models. The general form of these models can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \quad [7]$$

Where Y_i represents the outcome variable (either reading or math code proficiency) for individual i, β_0 is the intercept, β_k represents the coefficient for each independent variable X_k, and ε_i is the error term.

I estimated four specific model specifications for each outcome (reading and math):

$$Y_i = \beta_0 + \beta_1 \log(TUITION_i) + \beta_2 AGE_i + \varepsilon_i \quad [7(i)]$$

$$Y_i = \beta_0 + \beta_1 \log(TUITION_i) + \beta_2 AGE_i + \beta_3 MALE_i + \ldots + \beta_k X_{ki} + \varepsilon_i \quad [7(ii)]$$

$$Y_i = \beta_0 + \beta_1 \log(TUITION_i) + \beta_2 MALE_i + \beta_3 (\log(TUITION_i) \times MALE_i) + \beta_4 AGE_i + \ldots + \beta_k X_{ki} + \varepsilon_i \quad [7(iii)]$$

$$Y_i = \beta_0 + \beta_1 \log(TUITION_i) + \beta_2 (\log(TUITION_i))^2 + \beta_3 AGE_i + \ldots + \beta_k X_{ki} + \varepsilon_i \quad [7(iv)]$$

In Equation 7 (i-iv), the key independent variable, log(TUITION_i), is the natural logarithm of tuition amount, to account for the potentially non-linear relationship between tuition spending and academic performance, and to reduce the influence of extreme values. AGE_i represents the child's age, MALE_i is a dummy variable for gender, and X_ki represents other control variables.

All models were estimated using the subsample of students who pay tuition, allowing for an analysis of how variations in tuition amount relate to academic outcomes among those who invest in additional education.

I structured my analysis to examine these relationships across the entire sample, as well as separately for primary and secondary school students. This stratified approach allowed me to capture potential differences in the impact of tuition amount at different educational stages.

Throughout the analysis, I have strived to employ a comprehensive and methodologically sound approach. By utilizing multiple regression techniques, accounting for interaction effects, and addressing potential sources of endogeneity, I aimed to provide a nuanced and reliable analysis of the factors influencing educational outcomes and in particular the contribution of shadow education to learning outcomes.

5.0 Results and Findings:

Table 1: Descriptive Statistics

	observations	mean	sd	min	max
read_code	370,250	3.63	1.48	1	5
math_code	369,840	3.33	1.28	1	5
tuition_dummy	417,477	0.23	0.42	0.00	1
tuition_amount	92,368	272.6	309.91	1	5,000.00
log_tuition_amount	414,014	1.19	2.24	0.00	8.52
child_age	443,333	10.52	3.16	5	16
male	438,710	0.52	0.50	0.00	1
total_member	442,271	6.48	3.09	1	77
father_edu	413,257	0.74	0.44	0.00	1
mother_edu	432,855	0.55	0.50	0.00	1
wealth_index	414,697	2.4	1.22	0.00	6
wealth_index_pca1	414,697	0.01	1.32	-2.3	4.43
wealth_index_pca2	414,697	0.02	1	-5.13	1.25
hh_electricity	440,790	0.83	0.38	0.00	1
hh_4wheeler	434,033	0.08	0.27	0.00	1
hh_2wheeler	438,936	0.35	0.48	0.00	1
hh_has_newspaper	437,796	0.10	0.30	0.00	1
hh_has_mobile	436,561	0.80	0.40	0.00	1
hh_reading	436,568	0.25	0.43	0.00	1
in_school	443,468	1	0.00	1	1
in_primary_school	443,468	0.52	0.50	0.00	1
in_secondary_school	443,468	0.44	0.50	0.00	1
govt_school	292,991	1	0.00	1	1
private_school	146,732	1	0.00	1	1
madarsa_school	3,252	1	0.00	1	1
other_school	493	1	0.00	1	1

Source: ASER Dataset 2016

In my study, I analysed the ASER 2016 dataset of school-age children in India. I have narrowed my sample size from 639,752 to 443,468 to include only children who are in school, with 52% in primary school and 44% in secondary school. The majority attend government schools (n = 292,991), followed by private schools (n = 146,732), with a small number in madrasas (n = 3,252) and other types of schools (n = 493).

The educational outcomes, as measured by reading and mathematics codes, reveal interesting patterns. The mean reading code is 3.63 on a scale of 1 to 5, while the mean mathematics code is slightly lower at 3.33. These figures suggest that, on average, students in the sample demonstrate moderate proficiency in both subjects, with slightly better performance in reading compared to mathematics.

A key focus of my study is the prevalence and impact of private tutoring. I found that 23% of students in the sample receive additional tuition. Among those who do, the average monthly expenditure on tuition is 272.60 rupees, with a wide range from 1 to 5,000 rupees. To account for the skewed distribution of tuition costs, I also calculated the log of tuition amount where the mean is 1.19. This substantial investment in supplementary education by some families underscores the perceived importance of education and potentially points to gaps in the formal education system.

The demographic characteristics of the sample provide important context. The average age of children in the study is 10.52 years, ranging from 5 to 16 years. The gender distribution is nearly balanced, with 52% male students. The average household size is 6.48 members, ranging from 1 to 77, indicating significant variability in family structures.

Parental education shows interesting patterns. From the dummy variables, I observed that 74% of fathers and 55% of mothers have some level of education. Although this might lack the nuances of higher-level education attainment, however, the blatant gender disparity in parental education levels may have implications for household dynamics and children’s educational outcomes.

In my study, I utilized a wealth index to capture the socioeconomic status of households. This index, ranging from 0 to 6, has a mean of 2.404, suggesting that the average household in my sample possesses between two and three wealth indicators. The index represents a sum of key assets or favourable household characteristics, with each positive attribute contributing one point. This approach provides an intuitive measure of relative wealth across the sample, allowing for easy categorization of households into low, medium, and high wealth groups. I also constructed two wealth indices using principal component analysis (PCA) to allow for a more nuanced analysis of socioeconomic status beyond individual asset ownership.

In terms of household amenities and assets, 83% of households have access to electricity. Vehicle ownership varies, with 8% of households owning a four-wheeler and 35% owning a two-wheeler. Notably, mobile phone penetration is high at 80%. However, only 10% of households report having a newspaper, and 25% have reading materials. These figures suggest a population with varying levels of economic resources and a potential disconnect between technological adoption and traditional literacy materials.

I included state-level dummy variables to capture regional variations across India. These variables provide valuable insights into the geographic distribution of my sample and allow for the analysis of state-specific effects on educational outcomes and tutoring practices. Let me elaborate on the state-wise representation in my dataset:

Uttar Pradesh has the highest representation in the sample, with 14.68% of observations. This is followed by Bihar at 9.27% and Madhya Pradesh at 8.95%. These three states together account for about one-third of the sample, reflecting their large populations and potentially indicating a focus on states with significant educational challenges.

Rajasthan 6.46% and Maharashtra 4.75% also have substantial representation. Tamil Nadu follows with 4.45%, and Assam with 4.16% of the sample.

Gujarat 3.45% and West Bengal 2.33%, have smaller but still significant representations. Himachal Pradesh 1.79%, and Kerala 1.48%, have the lowest representation among the specifically identified states. A substantial portion of the sample 38.22% is categorized as “Other States” where I have included all the other states.

The inclusion of these state dummies in my analysis will allow me to control for state-specific factors that might influence educational outcomes and tutoring practices. This could include differences in state educational policies, economic conditions, cultural attitudes towards education, or the prevalence of private schooling and tutoring.

5.1 Ordinary Least Square (OLS):

In Model 1, I conducted an OLS regression to examine the relationship between reading code proficiency, tuition status, and child age. This model explains approximately 36.9% of the variance in reading code scores (R-squared = 0.369), providing significant insights into the factors influencing reading proficiency.

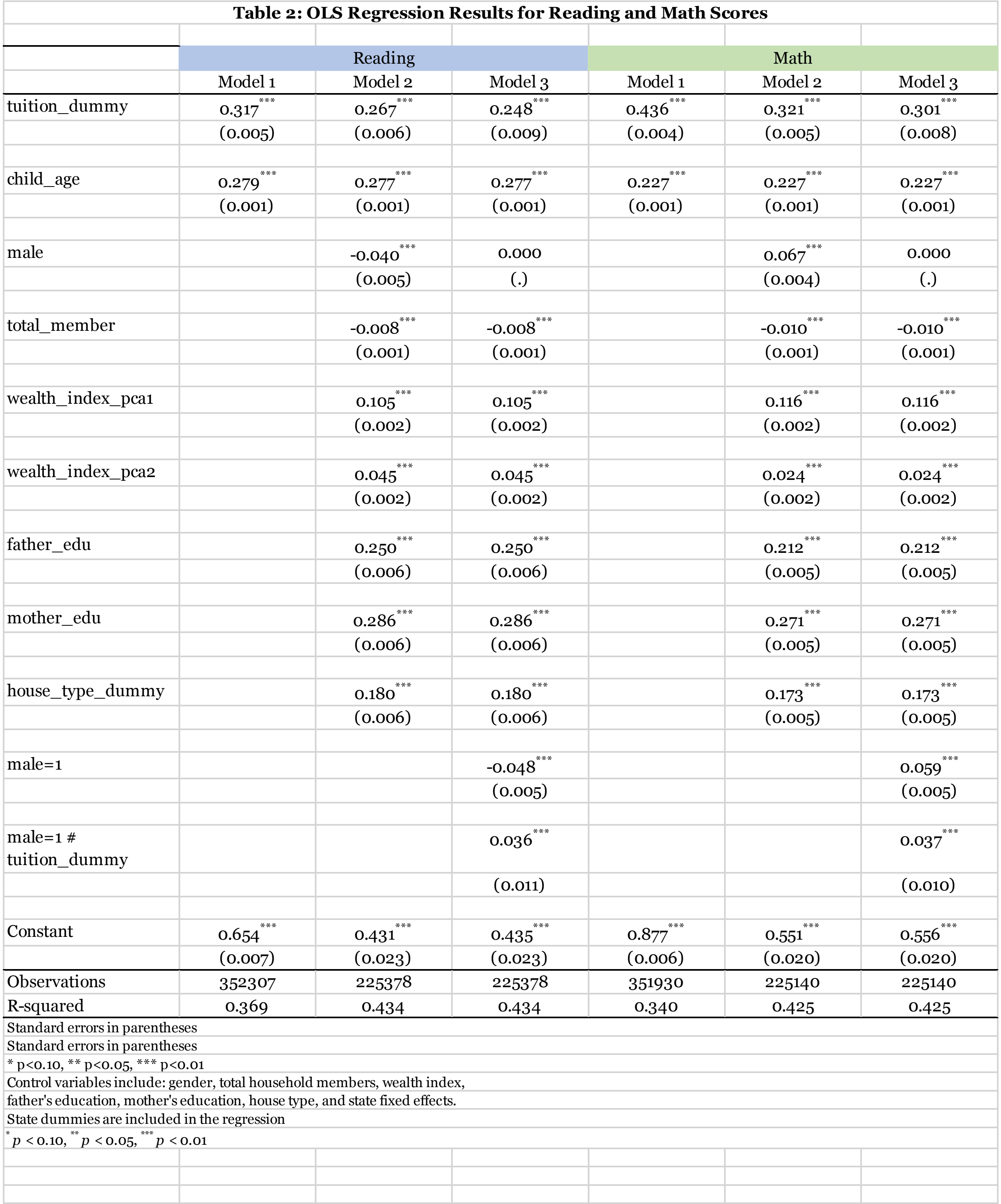

Table 2: OLS Regression Results for Reading and Math Scores

	Reading			Math		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
tuition_dummy	0.317***	0.267***	0.248***	0.436***	0.321***	0.301***
	(0.005)	(0.006)	(0.009)	(0.004)	(0.005)	(0.008)
child_age	0.279***	0.277***	0.277***	0.227***	0.227***	0.227***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
male		-0.040***	0.000		0.067***	0.000
		(0.005)	(.)		(0.004)	(.)
total_member		-0.008***	-0.008***		-0.010***	-0.010***
		(0.001)	(0.001)		(0.001)	(0.001)
wealth_index_pca1		0.105***	0.105***		0.116***	0.116***
		(0.002)	(0.002)		(0.002)	(0.002)
wealth_index_pca2		0.045***	0.045***		0.024***	0.024***
		(0.002)	(0.002)		(0.002)	(0.002)
father_edu		0.250***	0.250***		0.212***	0.212***
		(0.006)	(0.006)		(0.005)	(0.005)
mother_edu		0.286***	0.286***		0.271***	0.271***
		(0.006)	(0.006)		(0.005)	(0.005)
house_type_dummy		0.180***	0.180***		0.173***	0.173***
		(0.006)	(0.006)		(0.005)	(0.005)
male=1			-0.048***			0.059***
			(0.005)			(0.005)
male=1 # tuition_dummy			0.036***			0.037***
			(0.011)			(0.010)
Constant	0.654***	0.431***	0.435***	0.877***	0.551***	0.556***
	(0.007)	(0.023)	(0.023)	(0.006)	(0.020)	(0.020)
Observations	352307	225378	225378	351930	225140	225140
R-squared	0.369	0.434	0.434	0.340	0.425	0.425

Standard errors in parentheses
Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01
Control variables include: gender, total household members, wealth index,
father's education, mother's education, house type, and state fixed effects.
State dummies are included in the regression
* p < 0.10, ** p < 0.05, *** p < 0.01

The coefficient for the tuition dummy variable is 0.317, which is statistically significant at the 1% level. This indicates that children who receive tuition score, on average, 0.317

points higher on the reading code assessment compared to those who do not receive tuition, holding age constant. Child age demonstrates a strong positive relationship with reading code proficiency. The coefficient for child age is 0.279, which is also statistically significant at the 1% level. This suggests that, on average, for each year increase in a child's age, their reading code score increases by approximately 0.279 points, holding tuition status constant. This underscores the importance of age and cognitive development in reading skills.

In Model 2, I expanded the regression analysis to examine the relationship between reading code proficiency, tuition status, and various control variables including demographic factors and state-specific effects. This model explains approximately 43.4% of the variance in reading code scores, a substantial improvement from Model 1, indicating a better fit and more comprehensive explanation of factors influencing reading proficiency.

The coefficient for the tuition dummy variable is 0.267, which is statistically significant at the 1% level. This indicates that children who receive tuition score, on average, 0.267 points higher on the reading code assessment compared to those who do not receive tuition, holding other factors constant. While still substantial, this effect is slightly lower than in Model 1, suggesting that some of the apparent tuition effect was actually due to other factors now controlled for.

Gender shows a significant effect, with males scoring 0.04 points lower than females on average. Child age remains a strong predictor, with each year increase associated with a 0.277 point increase in reading code score, very similar to Model 1. Parental education shows substantial positive effects, with father's education associated with a 0.250 point increase and mother's education with a 0.286 point increase in reading code score for each level of education. This highlights the importance of parental background in children's reading achievement.

Wealth indices and housing type are positively associated with reading proficiency, indicating that socioeconomic factors play a role in reading achievement. Specifically, a one-unit increase in the primary wealth index (component 1) is associated with a 0.105 point increase in reading code score.

State-specific effects show considerable variation. For instance, children in Himachal Pradesh score 0.342 points higher than the reference state (Kerala), while those in Bihar score 0.293 points lower, all else being equal.

In Model 3, I expanded the regression analysis to include an interaction term of tuition dummy and gender. This model explains approximately 43.4% of the variance in reading code scores, similar to Model 2, indicating a consistent fit.

The coefficient for the tuition dummy variable is 0.248, which is statistically significant at the 1% level. This indicates that for females (the reference category for gender), receiving tuition is associated with a 0.248 point increase in reading code scores on average, holding other factors constant. The coefficient for being male is -0.048, suggesting that males score 0.048 points lower than females on average in reading code proficiency, all else being equal.

Interestingly, the interaction term between tuition and being male is positive (0.036) and significant at the 1% level. This suggests that the effect of tuition on reading code proficiency is more pronounced for males. Specifically, for males, the total effect of tuition is the sum of the main effect and the interaction effect (0.248 + 0.036 = 0.284), which is higher than for females.

In Model 4, I conducted an OLS regression to examine the relationship between math code proficiency, tuition status, and child age. This model explains approximately 34.0% of the variance in math code scores. The coefficient for the tuition dummy variable is 0.436, which is statistically significant at the 1% level. This indicates that children who receive tuition score, on average, 0.436 points higher on the math code assessment compared to those who do not receive tuition, holding age constant. This represents a substantial increase in math proficiency associated with tuition.

Child age demonstrates a strong positive relationship with math code proficiency. The coefficient suggests that, on average, for each year increase in a child's age, their math code score increases by approximately 0.227 points, holding tuition status constant.

In Model 5, I expanded the regression analysis to include other control variables. This model explains approximately 42.5% of the variance in math code scores, a substantial improvement from Model 4, indicating a better fit and more comprehensive explanation of factors influencing mathematical proficiency.

The coefficient for the tuition dummy variable is 0.321, which is statistically significant at the 1% level. This indicates that children who receive tuition score, on average, 0.321 points higher on the math code assessment compared to those who do not receive tuition, holding other factors constant. While still substantial, this effect is lower than in Model

4, suggesting that some of the apparent tuition effect was actually due to other factors now controlled for.

Gender shows a significant effect, with males scoring 0.067 points higher than females on average. Child age remains a strong predictor, with each year increase associated with a 0.227 point increase in math code score, very similar to Model 4.

Parental education shows substantial positive effects, with father's education associated with a 0.212 point increase and mother's education with a 0.271 point increase in math code score for each level of education. This highlights the importance of parental background in children's mathematical achievement.

Wealth indices and housing type are positively associated with math proficiency, indicating that socioeconomic factors play a role in mathematical achievement. Specifically, a one-unit increase in the primary wealth index (component 1) is associated with a 0.116 point increase in math code score.

State-specific effects show considerable variation. For instance, children in Himachal Pradesh score 0.414 points higher than the reference state (Kerala), while those in Maharashtra score 0.327 points lower, all else being equal.

Model 6 explains approximately 42.5% of the variance in math code scores, similar to Model 5, indicating a consistent fit. The coefficient for the tuition dummy variable is 0.301, which is statistically significant at the 1% level. This indicates that for females (the reference category for gender), receiving tuition is associated with a 0.301-point increase in math code scores on average, holding other factors constant.

The coefficient for being male is 0.059, suggesting that males score 0.059 points higher than females on average in math code proficiency, all else being equal.

Interestingly, the interaction term between tuition and being male is positive (0.037) and significant at the 1% level. This suggests that the effect of tuition on math code proficiency is more pronounced for males. Specifically, for males, the total effect of tuition is the sum of the main effect and the interaction effect (0.301 + 0.037 = 0.338), which is higher than for females.

Child age remains a strong predictor, with each year increase associated with a 0.227 point increase in math code score, consistent with previous models. Parental education maintains substantial positive effects, with father's education associated with a 0.212 point increase and mother's education with a 0.271 point increase in math code score for each level of education.

Wealth indices and housing type continue to be positively associated with math proficiency. A one-unit increase in the primary wealth index (component 1) is associated with a 0.116-point increase in math code score. State-specific effects continue to show considerable variation. For instance, children in Himachal Pradesh score 0.414 points higher than the reference state (Kerala), while those in Maharashtra score 0.327 points lower, all else being equal.

5.2 OLS across school types:

In my analysis of these eight regression models, I focused into the determinants of educational outcomes in secondary and primary schools, both government and private.

Table 3: OLS Regression Results for Reading and Math Scores in different School Types

	Reading				Math			
	Secondary Govt	Secondary Private	Primary Govt	Primary Private	Secondary Govt	Secondary Private	Primary Govt	Primary Private
tuition_dummy	0.239***	0.050***	0.421***	0.153***	0.358***	0.156***	0.372***	0.219***
	(0.010)	(0.010)	(0.012)	(0.014)	(0.010)	(0.012)	(0.010)	(0.012)
male	-0.030***	-0.059***	-0.046***	-0.099***	0.088***	0.033***	0.045***	0.022**
	(0.007)	(0.008)	(0.009)	(0.012)	(0.007)	(0.010)	(0.007)	(0.010)
total_member	-0.007***	0	-0.016***	-0.006***	-0.009***	-0.005***	-0.015***	-0.008***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)
wealth_index_pca1	0.076***	0.038***	0.114***	0.105***	0.090***	0.078***	0.110***	0.106***
	(0.003)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.003)	(0.004)
wealth_index_pca2	0.044***	0.008	0.025***	0.042***	0.023***	0.016***	0.007*	0.044***
	(0.004)	(0.005)	(0.005)	(0.006)	(0.004)	(0.006)	(0.004)	(0.005)
child_age	0.117***	0.067***	0.321***	0.319***	0.101***	0.064***	0.267***	0.264***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)
father_edu	0.197***	0.124***	0.255***	0.298***	0.182***	0.153***	0.187***	0.248***
	(0.009)	(0.013)	(0.010)	(0.019)	(0.009)	(0.015)	(0.008)	(0.015)
mother_edu	0.187***	0.125***	0.354***	0.316***	0.224***	0.190***	0.281***	0.256***
	(0.008)	(0.011)	(0.010)	(0.016)	(0.009)	(0.012)	(0.008)	(0.012)
house_type_dummy	0.102***	0.075***	0.199***	0.177***	0.120***	0.111***	0.144***	0.164***
	(0.009)	(0.012)	(0.010)	(0.016)	(0.009)	(0.014)	(0.008)	(0.013)
Constant	2.655***	3.551***	0.262***	0.441***	2.228***	3.007***	0.331***	0.554***
	(0.045)	(0.046)	(0.058)	(0.055)	(0.045)	(0.054)	(0.047)	(0.045)
Observations	67765	30879	77054	41169	67691	30844	76978	41129
R-squared	0.133	0.063	0.318	0.267	0.170	0.131	0.330	0.303

Standard errors in parentheses

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Control variables include: gender, total household members, wealth index,
father's education, mother's education, house type, and state fixed effects.

State dummies are included in the regression

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

The coefficient on the tuition dummy is consistently positive and statistically significant at the 1% level across all models. Its economic significance is particularly notable in government schools, with coefficients ranging from 0.239 to 0.421 for reading scores and 0.358 to 0.372 for math scores. This suggests that private tuition is associated with a substantial increase in test scores, potentially offsetting some of the disadvantages faced by students in government schools.

Gender effects are statistically significant but show varying economic impacts. The male coefficient is negative for reading scores (ranging from -0.030 to -0.099) and positive for math scores (ranging from 0.022 to 0.088), all significant at the 1% level except for primary private math scores (5% level). This indicates a gender gap in subject performance, with the gap being more pronounced in private schools for reading.

Socioeconomic indicators, including parental education and wealth proxies (house type dummy and wealth index components), show consistently positive and statistically significant coefficients. The economic impact of these factors is substantial, with parental education coefficients ranging from 0.124 to 0.354. This underscores the strong influence of family background on educational outcomes.

Age effects are positive and statistically significant at the 1% level across all models. The coefficients are larger for primary schools (ranging from 0.264 to 0.321) compared to secondary schools (0.064 to 0.117), suggesting a more pronounced impact of age on academic performance in earlier years of schooling.

State fixed effects reveal significant regional disparities. Using Kerala as the base state, I found statistically significant differences in most other states. For instance, Himachal Pradesh consistently shows positive coefficients (ranging from 0.146 to 0.461), while states like Uttar Pradesh and Bihar often display negative coefficients (as low as -0.838 for Bihar in primary government reading scores). These state effects persist even after controlling for individual and household characteristics, indicating substantial unexplained regional variation in educational outcomes.

The models' explanatory power, as indicated by R-squared values, is generally higher for primary schools (ranging from 0.267 to 0.330) compared to secondary schools (0.063 to 0.170). This suggests that the included variables explain a larger proportion of the variance in primary school outcomes, while secondary school performance may be influenced by additional factors not captured in these models.

5.3 Marginal Effect:

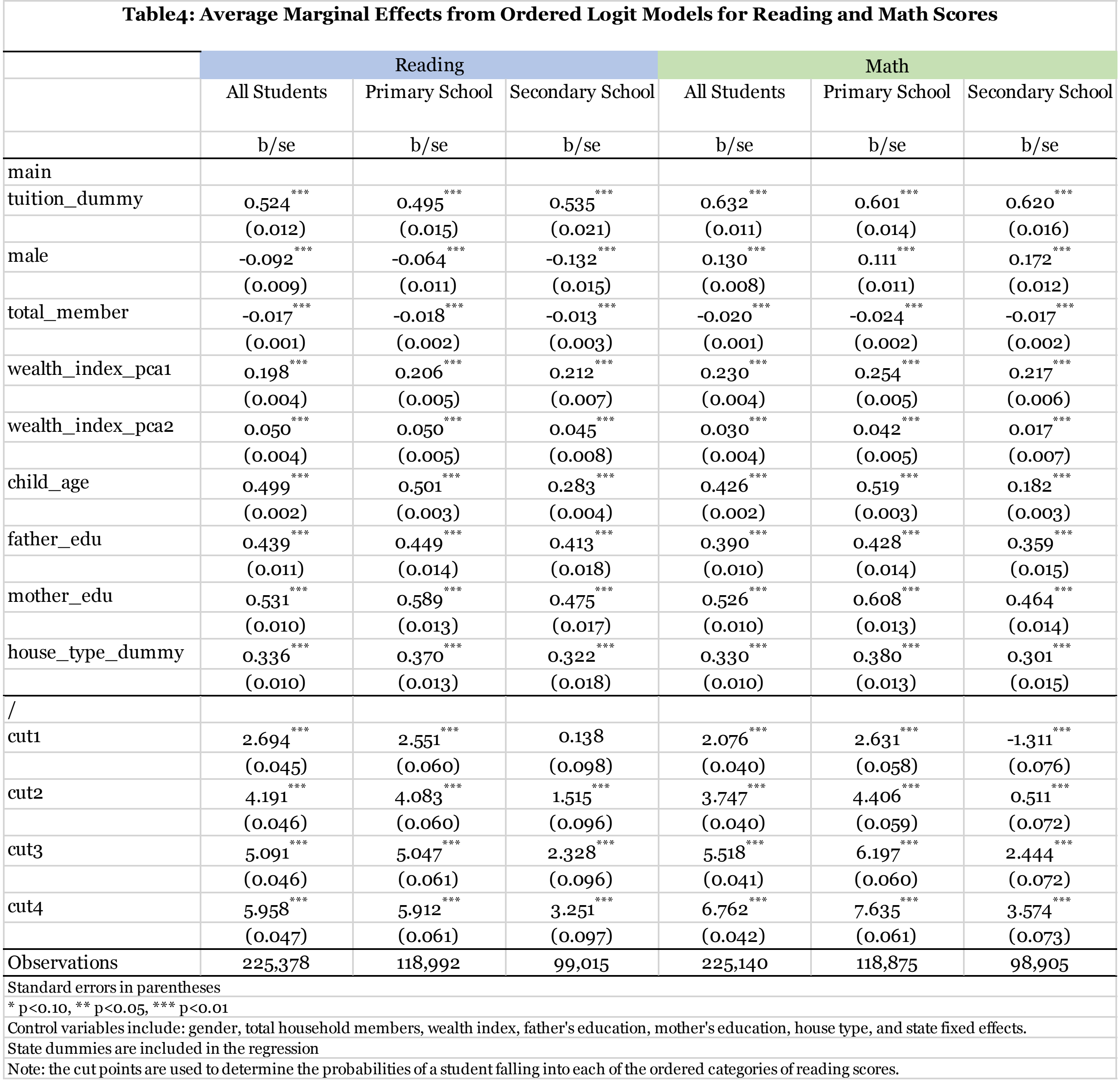

Table4: Average Marginal Effects from Ordered Logit Models for Reading and Math Scores

	Reading			Math		
	All Students	Primary School	Secondary School	All Students	Primary School	Secondary School
	b/se	b/se	b/se	b/se	b/se	b/se
main						
tuition_dummy	0.524***	0.495***	0.535***	0.632***	0.601***	0.620***
	(0.012)	(0.015)	(0.021)	(0.011)	(0.014)	(0.016)
male	-0.092***	-0.064***	-0.132***	0.130***	0.111***	0.172***
	(0.009)	(0.011)	(0.015)	(0.008)	(0.011)	(0.012)
total_member	-0.017***	-0.018***	-0.013***	-0.020***	-0.024***	-0.017***
	(0.001)	(0.002)	(0.003)	(0.001)	(0.002)	(0.002)
wealth_index_pca1	0.198***	0.206***	0.212***	0.230***	0.254***	0.217***
	(0.004)	(0.005)	(0.007)	(0.004)	(0.005)	(0.006)
wealth_index_pca2	0.050***	0.050***	0.045***	0.030***	0.042***	0.017***
	(0.004)	(0.005)	(0.008)	(0.004)	(0.005)	(0.007)
child_age	0.499***	0.501***	0.283***	0.426***	0.519***	0.182***
	(0.002)	(0.003)	(0.004)	(0.002)	(0.003)	(0.003)
father_edu	0.439***	0.449***	0.413***	0.390***	0.428***	0.359***
	(0.011)	(0.014)	(0.018)	(0.010)	(0.014)	(0.015)
mother_edu	0.531***	0.589***	0.475***	0.526***	0.608***	0.464***
	(0.010)	(0.013)	(0.017)	(0.010)	(0.013)	(0.014)
house_type_dummy	0.336***	0.370***	0.322***	0.330***	0.380***	0.301***
	(0.010)	(0.013)	(0.018)	(0.010)	(0.013)	(0.015)
/						
cut1	2.694***	2.551***	0.138	2.076***	2.631***	-1.311***
	(0.045)	(0.060)	(0.098)	(0.040)	(0.058)	(0.076)
cut2	4.191***	4.083***	1.515***	3.747***	4.406***	0.511***
	(0.046)	(0.060)	(0.096)	(0.040)	(0.059)	(0.072)
cut3	5.091***	5.047***	2.328***	5.518***	6.197***	2.444***
	(0.046)	(0.061)	(0.096)	(0.041)	(0.060)	(0.072)
cut4	5.958***	5.912***	3.251***	6.762***	7.635***	3.574***
	(0.047)	(0.061)	(0.097)	(0.042)	(0.061)	(0.073)
Observations	225,378	118,992	99,015	225,140	118,875	98,905

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Control variables include: gender, total household members, wealth index, father's education, mother's education, house type, and state fixed effects.

State dummies are included in the regression

Note: the cut points are used to determine the probabilities of a student falling into each of the ordered categories of reading scores.

In this comprehensive analysis, I examined the factors influencing reading and mathematical proficiency among primary and secondary school students, based on the marginal effects derived from multinomial logistic regression models. The data presents a nuanced picture of educational outcomes, highlighting the complex interplay of socioeconomic, demographic, and individual factors that shape students' academic achievements.

At the outset, it is crucial to note that the models employ a five-level coding system for both reading and mathematical proficiency, allowing for a granular analysis of the impacts across different skill levels. The independent variables consistently include factors such as tuition attendance, gender, family size, wealth indices, child age, parental education, and housing type. This consistency across models facilitates a comparative analysis of how these factors' influences evolve from primary to secondary education and between reading and mathematical skills.

One of the most striking findings across all models is the substantial impact of tuition or formal schooling on both reading and mathematical proficiency. In the primary school reading model, attending school or receiving tuition decreases the probability of being in the lowest reading level by 6 percentage points and increases the probability of being in the highest level by 9 percentage points. This effect is even more pronounced in secondary school, where tuition reduces the likelihood of being in the lowest reading level by 9.9 percentage points and increases the chances of being in the highest level by 13.3 percentage points. The impact on mathematical skills follows a similar pattern, with tuition increasing the probability of being in the highest math level by 5 percentage points in primary school and an impressive 13.3 percentage points in secondary school.

These results underscore the critical role of formal education in developing both literacy and numeracy skills. The increasing magnitude of the effect from primary to secondary education suggests that the benefits of schooling are cumulative, with continued education yielding progressively larger returns in terms of skill development. From an economic perspective, this highlights the importance of investments in education as a means of human capital development. The substantial gains in proficiency associated with schooling indicate that policies aimed at Increasing school attendance and reducing dropout rates could yield significant societal benefits in terms of improved literacy and numeracy.

Gender differences in educational outcomes present an intriguing pattern that varies between subjects and educational levels. In reading proficiency, girls show a slight advantage over boys, which becomes more pronounced in secondary education. For primary school students, being male decreases the probability of being in the highest reading level by 1.4 percentage points, while in secondary school, this gap widens to 2.2 percentage points. Conversely, in mathematics, boys exhibit a small advantage that also increases from primary to secondary education. In primary school, being male increases the probability of being in the highest math level by 0.9 percentage points, which grows to 3.7 percentage points in secondary school.

These gender disparities, while relatively small, are statistically significant and point to persistent gender-based differences in educational outcomes. The widening of these gaps from primary to secondary education is particularly concerning, as it suggests that initial small differences may compound over time. From a policy perspective, these findings call for targeted interventions to address gender-specific challenges in education. For reading, efforts may be needed to enhance boys' engagement with literacy, while in mathematics, strategies to boost girls' confidence and participation could be beneficial. The economic implications of these gender gaps are significant, as they may translate into occupational segregation and wage disparities in the labour market if left unaddressed.

Family size emerges as a consistent, albeit small, negative influence on both reading and mathematical proficiency. In primary education, each additional family member decreases the probability of being in the highest reading level by 0.26 percentage points and the highest math level by 0.2 percentage points. These effects, while slightly diminished, persist into secondary education. The negative impact of larger family size on educational outcomes likely reflects resource constraints within households. In economic terms, this suggests that there may be a trade-off between quantity and quality of children, as theorized by Gary Becker. Policies that aim to support larger families, such as targeted educational subsidies or family planning initiatives, could help mitigate these negative effects.

The influence of socioeconomic status, as measured by wealth indices and housing type, is substantial and consistent across all models. Higher wealth is strongly associated with better outcomes in both reading and mathematics, with effects that persist from primary to secondary education. For instance, in primary education, an increase in the first wealth index raises the probability of being in the highest reading level by 3.1 percentage points and the highest math level by 2.1 percentage points. These effects are even larger in secondary education. Similarly, better housing conditions, likely indicative of higher socioeconomic status, are associated with improved outcomes in both subjects and at both educational levels.

These findings highlight the pervasive nature of socioeconomic inequality in educational outcomes. The persistence and, in some cases, widening of these effects from primary to secondary education suggests that initial socioeconomic disadvantages may compound over time, leading to increasing disparities in human capital accumulation. From an economic perspective, this represents a significant inefficiency, as it implies that talented individuals from disadvantaged backgrounds may not be realizing their full potential. Policies aimed at reducing socioeconomic barriers to education, such as targeted financial assistance, provision of educational resources, or community-based interventions, could yield substantial returns in terms of improved educational outcomes and, ultimately, economic productivity.

The strong positive effect of age on both reading and mathematical proficiency across all models reflects the natural progression of skill development as children mature and advance through their education. In primary education, each year increase in age is associated with a 4.8 percentage point increase in the probability of being in the highest reading level and a 4.3 percentage point increase for mathematics. While still significant, these effects are slightly smaller in secondary education, possibly indicating a more gradual improvement in skills at higher educational levels.

The age effects underscore the cumulative nature of skill development in education. From a policy perspective, this highlights the importance of early interventions to ensure that children start their educational journey on a strong footing. Additionally, the continued significant effects in secondary education suggest that there are ongoing opportunities for skill development, emphasizing the value of sustained educational investments throughout a student’s academic career.

Parental education, particularly maternal education, emerges as a crucial factor in determining children’s educational outcomes. In primary education, an increase in mother’s education is associated with an 8 percentage point increase in the probability of being in the highest reading level and a 5.1 percentage point increase for mathematics. These effects are even larger in secondary education, with maternal education increasing the likelihood of being in the highest reading level by 8 percentage points and the highest math level by 9.9 percentage points. Father’s education also shows significant positive effects, though generally smaller in magnitude than maternal education.

The strong influence of parental education, especially maternal education, on children's academic performance highlights the intergenerational transmission of human capital. This finding has important implications for understanding the persistence of educational inequalities across generations. From an economic perspective, it suggests that investments in education can have multiplier effects, as educated parents are more likely to raise educated children. Policies that focus on adult education and parental involvement in children’s education could, therefore, yield long-term benefits in terms of improved educational outcomes across generations.

The consistency and magnitude of these effects across different subjects and educational levels underscore the robustness of the findings. The use of multinomial logistic regression allows for a nuanced understanding of how various factors affect the probability of students being at different proficiency levels, providing valuable insights beyond simple binary outcomes.

However, it is important to note some limitations of this analysis. While the models control for a range of factors, there may be unobserved variables that influence educational outcomes. Additionally, the cross-sectional nature of the data limits our ability to make causal inferences. Longitudinal studies would be valuable in further elucidating the causal mechanisms underlying these associations.

In conclusion, this analysis reveals a complex landscape of factors influencing educational outcomes in reading and mathematics across primary and secondary education. The persistent and often increasing effects of socioeconomic factors, parental education, and gender from primary to secondary education highlight the cumulative nature of educational advantages and disadvantages. These findings have significant implications for educational policy and economic development strategies.

From an economic perspective, the results underscore the importance of education as a means of human capital development and highlight several areas where targeted interventions could yield substantial returns. The strong positive effects of formal education suggest that investments in expanding access to quality schooling could significantly boost literacy and numeracy skills. The persistent socioeconomic gradients in educational outcomes point to the need for policies that address broader social inequalities and provide additional support to disadvantaged students.

The gender differences observed, particularly their widening from primary to secondary education, call for targeted approaches to ensure gender equity in educational outcomes. This is crucial not only from an equity standpoint but also for maximizing human capital development and economic productivity.

The significant influence of parental education, especially maternal education, highlights the potential for intergenerational effects of educational investments. Policies that support adult education and parental involvement in children's schooling could have far-reaching impacts on educational outcomes across generations.

Finally, the consistent negative effect of family size on educational outcomes, albeit small, suggests that family planning policies and targeted support for larger families could play a role in improving educational attainment.

In sum, these findings provide a rich empirical basis for developing nuanced, evidence-based policies to improve educational outcomes. By addressing the multifaceted determinants of academic achievement, policymakers can work towards creating more equitable and effective educational systems, ultimately contributing to broader goals of economic development and social mobility.

5.4 Family Fixed Effect:

In my analysis of these family fixed effects models, I examined the impact of private tuition on reading and math scores while controlling for unobserved family-level characteristics. This approach allows me to isolate the effect of tuition within families, effectively controlling for factors that are constant across siblings.

Table5: Family Fixed Effects Models for Reading and Math Scores

	Reading			Math		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
tuition_dummy	0.725***	0.254***	0.231***	0.714***	0.314***	0.294***
	-0.014	-0.011	-0.012	-0.012	-0.009	-0.011
child_age		0.285***	0.285***		0.236***	0.236***
		-0.001	-0.001		-0.001	-0.001
male		-0.033***	0		0.063***	0
		-0.005	(.)		-0.004	(.)
male=0			0			0
			(.)			(.)
male=1			-0.042***			0.055***
			-0.006			-0.005
male=0 # tuition_dummy			0			0
			(.)			(.)
male=1 # tuition_dummy			0.043***			0.036***
			-0.012			-0.01
Constant	3.486***	0.626***	0.631***	3.189***	0.775***	0.779***
	-0.004	-0.01	-0.01	-0.003	-0.009	-0.009
Observations	352389	348663	348663	352012	348284	348284
R-squared within	0.019	0.419	0.419	0.026	0.404	0.404
R-squared between	0.019	0.369	0.369	0.035	0.33	0.33
R-squared overall	0.018	0.368	0.368	0.034	0.338	0.338
Standard errors in parentheses						
Standard errors in parentheses						
="* p<0.10	** p<0.05	*** p<0.01"				
Control variables in Models 2 and 3 include: gender and age						
="* p<0.10	** p<0.05	*** p<0.01"				

In Model 1, which includes only the tuition dummy, I found a large and statistically significant effect of tuition on both reading and math scores. The coefficients (0.725 for reading and 0.714 for math, both significant at the 1% level) suggest that receiving tuition is associated with substantial improvements in test scores. However, these estimates are likely biased due to omitted variables.

Model 2 introduces controls for age and gender, leading to a notable reduction in the tuition effect. The coefficients for the tuition dummy decrease to 0.254 for reading and 0.314 for math, remaining statistically significant at the 1% level. This reduction indicates that part of the tuition effect in Model 1 was capturing the influence of age and gender differences.

The age variable shows a strong positive effect (0.285 for reading and 0.236 for math, significant at 1%), consistent with expectations that older children perform better. The gender effect (male dummy) is negative for reading (-0.033) and positive for math (0.063), both significant at 1%, revealing a gender gap in subject performance.

Model 3 introduces an interaction between gender and tuition. The main effect of tuition remains similar to Model 2 (0.231 for reading and 0.294 for math, significant at 1%). Interestingly, the interaction term is positive and significant for both subjects (0.043 for reading and 0.036 for math, significant at 1%). This suggests that boys benefit more from tuition than girls, potentially exacerbating gender disparities in educational outcomes.

The R-squared values provide insights into the models' explanatory power. The within-family R-squared increases substantially from Model 1 (0.019 for reading, 0.026 for math) to Models 2 and 3 (around 0.41 for reading, 0.40 for math). This indicates that age and gender explain a considerable portion of within-family variation in test scores.

The between-family and overall R-squared values are lower than the within-family values, suggesting that unobserved family characteristics play a significant role in explaining differences in test scores across families. This underscores the importance of the family fixed effects approach in controlling for these unobserved factors.

5.5 OLS with Tuition Amount:

Table 6: OLS Regression Results for Reading and Math Scores for log of Tuition Amount

	Reading				Math			
	Secondary Govt	Secondary Private	Primary Govt	Primary Private	Secondary Govt	Secondary Private	Primary Govt	Primary Private
log_tuition_amount	0.090^{***}	0.052^{***}	0.156^{***}	0.131^{***}	0.104^{***}	0.090^{***}	0.156^{***}	0.143^{***}
	(0.010)	(0.012)	(0.018)	(0.019)	(0.011)	(0.014)	(0.015)	(0.015)
male	-0.015	-0.055^{***}	-0.016	-0.104^{***}	0.105^{***}	0.025	0.092^{***}	-0.01
	(0.012)	(0.016)	(0.021)	(0.024)	(0.013)	(0.019)	(0.018)	(0.019)
total_member	-0.002	0.002	-0.010^{***}	-0.008^{**}	-0.004^{**}	-0.003	-0.010^{***}	-0.007^{**}
	(0.002)	(0.002)	(0.003)	(0.004)	(0.002)	(0.003)	(0.003)	(0.003)
wealth_index_pca1	0.031^{***}	0.023^{***}	0.093^{***}	0.052^{***}	0.043^{***}	0.056^{***}	0.087^{***}	0.073^{***}
	(0.005)	(0.007)	(0.010)	(0.010)	(0.006)	(0.008)	(0.008)	(0.009)
wealth_index_pca2	0.012^{*}	0.004	0.010	0.017	0.014^{*}	0.032^{***}	0.010	0.045^{***}
	(0.007)	(0.009)	(0.012)	(0.014)	(0.007)	(0.011)	(0.010)	(0.011)
child_age	0.079^{***}	0.056^{***}	0.335^{***}	0.298^{***}	0.068^{***}	0.054^{***}	0.307^{***}	0.258^{***}
	(0.004)	(0.005)	(0.006)	(0.006)	(0.004)	(0.005)	(0.005)	(0.005)
father_edu	0.170^{***}	0.093^{***}	0.243^{***}	0.239^{***}	0.182^{***}	0.061^{*}	0.196^{***}	0.225^{***}
	(0.016)	(0.028)	(0.028)	(0.044)	(0.018)	(0.033)	(0.024)	(0.036)
mother_edu	0.128^{***}	0.079^{***}	0.377^{***}	0.285^{***}	0.140^{***}	0.146^{***}	0.304^{***}	0.249^{***}
	(0.014)	(0.021)	(0.025)	(0.035)	(0.015)	(0.025)	(0.022)	(0.028)
house_type_dummy	0.074^{***}	0.067^{***}	0.221^{***}	0.130^{***}	0.102^{***}	0.091^{***}	0.199^{***}	0.157^{***}
	(0.014)	(0.023)	(0.024)	(0.033)	(0.015)	(0.027)	(0.020)	(0.027)
Constant	2.887^{***}	3.489^{***}	-0.540^{***}	0.112	2.510^{***}	2.919^{***}	-0.642^{***}	-0.068
	(0.081)	(0.099)	(0.176)	(0.156)	(0.090)	(0.120)	(0.148)	(0.128)
Observations	17122	7327	13678	9709	17103	7322	13659	9702
R-squared	0.084	0.058	0.280	0.241	0.114	0.081	0.286	0.276
Adjusted R-squared	0.083	0.056	0.279	0.240	0.113	0.078	0.285	0.275
F-statistic	78.13	22.67	265.94	154.12	109.83	32.1	273.48	184.61

Standard errors in parentheses

Standard errors in parentheses

* p<0.10, ** p<0.05, *** p<0.01

Control variables include: gender, total household members, wealth index,
father's education, mother's education, house type, and state fixed effects.

State dummies are included in the regression

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

In my analysis of the OLS regression results for reading and math scores in relation to tuition amount across different school types, I observed several interesting patterns and relationships. The models explain varying portions of the variance in both reading and math scores, with R-squared values ranging from 0.058 to 0.286, indicating that the explanatory power differs considerably across school types and subjects.

The impact of tuition (log_tuition_amount) on both reading and math scores is consistently positive and statistically significant across all school types. For reading scores, a 100% increase in tuition is associated with the largest increase in primary government schools (0.156 points), followed by primary private schools (0.131 points), secondary government schools (0.090 points), and secondary private schools (0.052 points). For math scores, the effect is highest in primary government schools (0.156 points) and primary private schools (0.143 points), followed by secondary government schools (0.104 points) and secondary private schools (0.090 points). This suggests that tuition may be particularly beneficial in primary schools, possibly due to the foundational nature of learning at this stage.

Gender differences are apparent and vary across school types and subjects. For reading scores, males generally perform worse than females, with the largest gap in primary private schools where males score 10.4 points lower. The gender gap is also significant in secondary private schools (-5.5 points) but not statistically significant in government schools. For math scores, males tend to outperform females in government schools, with a 10.5 point higher score in secondary government schools and a 9.2 point higher score in primary government schools. However, this gender gap is not significant in private schools for math scores.

Age remains a strong predictor of performance in both subjects across all school types, with the effect being particularly pronounced in primary schools. For instance, in primary government schools, each year is associated with a 33.5% increase in reading scores and a 30.7% increase in math scores. In secondary schools, the effect is smaller but still significant, with each year associated with a 7.9% increase in reading scores and a 6.8% increase in math scores for government schools.

Parental education plays a significant role across all school types, with both father's and mother's education positively impacting scores. The effect of mother's education is particularly strong in primary government schools, where a one-unit increase in mother's education is associated with a 37.7% increase in reading scores and a 30.4% increase in math scores. Father's education has a stronger effect in secondary government schools, with a 17.0% increase in reading scores and an 18.2% increase in math scores for each unit increase.

The impact of household wealth, as indicated by the wealth index components and house type, is generally positive across all school types. For example, a one-unit increase in the first wealth index component is associated with a 2.3% to 9.3% increase in scores across various school types, suggesting that socioeconomic factors play a crucial role in educational outcomes.

The state-specific effects reveal considerable regional variation in educational outcomes across school types. For instance, students in Bihar perform 14.9% better than the reference state (Kerala) in secondary private schools for reading and 34.1% better for math. However, they perform significantly worse (50.3% lower) in primary government schools for reading. Rajasthan shows positive effects in secondary schools, with 14.0% and 19.1% higher reading scores in government and private schools respectively. These regional disparities could reflect differences in educational policies, resource allocation, or cultural factors affecting education across different states and school types.

.

6.0 Conclusion:

In conclusion, my research into the shadow education system in India reveals a complex landscape of educational investments and outcomes. Through rigorous econometric analysis of the ASER 2016 data, I have uncovered significant associations between private tutoring and improved learning outcomes in both reading and mathematics. These effects persist across various model specifications, including OLS regressions with extensive controls and family fixed effects models, suggesting a robust relationship between tutoring and academic performance.

The consistency of these findings across different empirical strategies lends credence to the notion that shadow education is indeed contributing to human capital accumulation in India. From an economic perspective, this indicates that households are making rational investments in their children’s education, seeking to augment the human capital provided by the formal school system. The persistence of tutoring effects even after controlling for socioeconomic status suggests that these investments are yielding returns over and above what might be expected from family background alone.

The analysis of tuition's impact on educational outcomes raises questions about equity in the education system. While we observe positive associations between tuition and both reading and math scores across school types, we haven't directly examined how tutoring receipt varies by socioeconomic status. However, without specific data on the socioeconomic distribution of tutoring, we should be cautious about drawing conclusions regarding the intergenerational transmission of advantage or long-term impacts on human capital distribution in the Indian economy. Further research would be needed to explore these broader societal implications.

From a policy standpoint, these findings present a dilemma. On one hand, the positive impacts of tutoring suggest that expanding access to such services could be a means of improving educational outcomes, particularly for disadvantaged students. On the other hand, the reliance on private tutoring to achieve satisfactory learning outcomes could be seen as a failure of the formal education system to meet the needs of all students. Policymakers may need to consider interventions that address both the demand for and supply of educational services, potentially through improvements in school quality, targeted support for disadvantaged students, or regulated integration of tutoring services into the formal education system.

The heterogeneity in tutoring effects across different subgroups, such as primary versus secondary students and boys versus girls, underscores the need for nuanced policy approaches. One-size-fits-all solutions are unlikely to be effective given the varying impacts of tutoring across these groups. Instead, targeted interventions that address the specific needs of different student populations may be necessary to maximize the benefits of educational investments.

This research highlights the importance of addressing endogeneity concerns in estimating the impacts of educational interventions. The family fixed effects models, in particular, provide a powerful tool for controlling unobserved family-level characteristics that might confound the relationship between tutoring and learning outcomes. The persistence of tutoring effects in these models provides strong evidence for a causal relationship, although further research using experimental or quasi-experimental designs could further strengthen these findings.

While this research provides valuable insights into the impact of shadow education in India, it is important to acknowledge its limitations. Firstly, the cross-sectional nature of the ASER data limits our ability to establish causal relationships definitively. Although our family fixed effects models address some endogeneity concerns, longitudinal data would allow for more robust causal inference. Secondly, our measure of tutoring is binary, which doesn't capture variations in the quality, intensity, or duration of tutoring received. A more nuanced measure could provide deeper insights into the effectiveness of different types of shadow education. Thirdly, while ASER data provides a comprehensive view of rural India, it doesn't include urban areas, potentially limiting the generalizability of our findings to the entire country. Additionally, our study focuses on reading and mathematics outcomes, which, while crucial, do not encompass the full spectrum of educational achievements. Finally, the potential selection bias in tutoring receipt - where more motivated students or parents might be more likely to seek tutoring - cannot be fully ruled out despite our econometric strategies. These limitations highlight the need for further research in this area, potentially using experimental designs or more detailed longitudinal data to address these gaps.

Looking forward, this research opens up several avenues for future inquiry. Longitudinal studies tracking students over time could provide insights into the long-term impacts of tutoring on educational and labor market outcomes. Additionally, more detailed data on the nature and quality of tutoring services could help unpack the mechanisms through which tutoring affects learning outcomes. Finally, comparative studies across different cultural and institutional contexts could shed light on how the shadow education phenomenon varies across different educational systems.

In sum, this research contributes to our understanding of the complex dynamics shaping educational outcomes in developing countries. By shedding light on the role of shadow education in India’s educational landscape, it provides valuable insights for policymakers seeking to improve educational quality and equity. As we continue to grapple with the challenges of human capital development in an increasingly knowledge-driven global economy, understanding and addressing the implications of shadow education will be crucial for creating more effective and equitable educational systems.

References

Roesgaard, Marie H. 2006. *Japanese Education and the Cram School Business: Functions, Challenges and Perspectives of the Juku*. Copenhagen: Nordic Institute of Asian Studies Press.

Seth, Michael J. 2002. *Education Fever: Society, Politics, and the Pursuit of Schooling in South Korea*. Honolulu: University of Hawai'i Press.

Sen, Amartya. 2010. "Primary Schooling in West Bengal." *Prospects: Quarterly Review of Comparative Education* 40 (3):311–320.

Sujatha, K. and P. Geetha Rani. 2011. *Management of Secondary Education in India*. New Delhi: Shipra and National University of Educational Planning and Administration.

Sujatha, K. (2014) Private tuition in India: trends and issues, *Revue Internationale d'éducation de Sèvres*. Available at: https://journals.openedition.org/ries/3913

Lee, Ji Yun. 2013. 'Private Tutoring and Its Impact on Students' Academic Achievement, Formal Schooling,and Educational Inequality in Korea', PhD, Columbia University. http://hdl.handle.net/10022/AC:P:20461

Brehm, William C. and Iveta Silova. 2014. 'Ethical Dilemmas in the Education Marketplace: Shadow Education, Political Philosophy and Social (In)justice in Cambodia', in Ian Macpherson, Susan Robertson, and Geoffrey Walford (eds)

Education, Privatisation and Social Justice: Case Studies from Africa,South Asia and South East Asia, Oxford: Symposium Books, pp. 159–178

Wadhwa, Wilima. 2015. 'Government vs Private Schools: Have Things Changed?', in *Annual Status of Education Report (Rural) 2014*, New Delhi: ASER Centre.

Banerji, Rukmini and Wilima Wadhwa. 2012. 'Every Child in School and Learning Well in India:Investigating the Implications of School Provision and Supplemental Help', *in India Infrastructure Report*,New Delhi: Routledge.

Desai, Sonalde B., Amaresh Dubey, Brij Lal Joshi, Mitali Sen, Abusaleh Sharif, and Reeve Vanneman. 2010.*Human Development in India: Challenges for a Society in Transition*, New Delhi: Oxford University Press.

Kumar, Krishna. 2012. 'Universities: Ours and Theirs', *The Hindu*, 9 August 2012.

Chakraborty, Sudhir. 2003. *Lekhapara Kore Je: A Collection of Bengali Essays on our Current Education World*. Kolkata: Dey's Publishing. (In Bengali)

Suraweera, A.V. 2011. *Dr. Kannangara's Free Education Proposals in Relation to the Subsequent Expansion of the Tuition Industry*. Dr. C.W.W. Kannangara Memorial Lecture 22, Mahargama, Sri Lanka: Department of Research and Development, National Institute of Education.

SHARMA, N. (2009) Public and Private Schools in Nepal: A Comparative

Perspective.

Kim, K. K. (2010). Educational equality. In C. J. Lee, S. Y. Kim, & D. Adams, (Eds.), *Sixty Years of Korean Education* (pp. 285–325). Seoul: Seoul National University Press.

Aslam, Monazza and Paul Atherton. 2011. "The "Shadow" Education Sector in India and Pakistan: The Determinants, Benefits and Equity Effects of Private Tutoring." Presentation at the UKFIET (United Kingdom Forum for International Education and Training) Conference, University of Oxford, 13–15 September.

Lee, Ji Yun. 2013. 'Private Tutoring and Its Impact on Students' Academic Achievement, Formal Schooling,and Educational Inequality in Korea', PhD, Columbia University. http://hdl.handle.net/10022/AC:P:20461

Majumdar, Manabi. 2014. 'The Shadow School System and New Class Divisions in India', TRG Poverty and Education Working Paper Series Paper 2, Max Weber Stiftung.

Majumdar, M., 2018. Access, success, and excess. In: R. Bhattacharya, M. Sharma & S. Vasudeva, eds. *Education and Inequality in India: A Classroom View*. 1st ed. Abingdon: Routledge, pp. 239-254. Available at: https://www.taylorfrancis.com/chapters/edit/10.4324/9781315107929-24/access-success-excess-manabi-majumdar [Accessed 30 Aug. 2024].

Dang, Hai-Anh and F. Halsey Rogers (2008). 'The Growing Phenomenon of Private Tutoring: Does it Deepen Human Capital, Widen Inequalities, or Waste Resources?' *World Bank Research Observer*, 23, 161200.

Kim, Gwang-Jo. 2001. “Education Policies and Reform in South Korea.” *In Secondary Education in Africa: Strategies for Renewal*. Washington, D.C.: World Bank.

Azam, M. (2016), Private Tutoring: Evidence from India. Rev Dev Econ, 20: 739-761. https://doi.org/10.1111/rode.12196

Statistics Korea. 2010. The Survey of Private Education Expenditures, 2009. Seoul: Auth

Appendix: Ethics Statement

In conducting this research on the impact of shadow education on educational outcomes in rural India, I have adhered to principles of transparency and replicability throughout my data analysis process. This statement outlines the key steps taken to ensure the integrity and reproducibility of my findings.

The data for this study was obtained from the Annual Status of Education Report (ASER) 2016, specifically the ASER 2016 Household Data file. This dataset is a large-scale household survey conducted in rural India, providing comprehensive information on children's educational status and household characteristics. I emailed the ASER data team to get hold of the dataset and they were kind enough to provide multiple years data. To prepare the data for analysis, I performed several data manipulation steps using Stata software.

I generated a series of dummy variables to represent key binary characteristics in the dataset. These included indicators for tuition status, parental education levels, household amenities, and child gender. For each of these variables, I used conditional statements to create binary indicators and subsequently cleaned the data by replacing any missing values with null entries. This process ensured the integrity of my data while preparing it for more complex analyses.

A central component of my data preparation was the construction of wealth indices to capture household economic status. I approached this in two ways. First, I created a simple additive wealth index by summing binary indicators for several household assets and amenities, including four-wheelers, two-wheelers, mobile phones, newspapers, electricity connections, and reading materials. This provided me with a straightforward measure of household wealth on a scale from 0 to 6. To complement this, I also employed Principal Component Analysis (PCA) on the same set of variables to generate two additional wealth indices. This dual approach to measuring wealth allowed me to test the robustness of my findings to different specifications of household economic status.

To focus on children currently attending school, I created variables to identify children currently in school and those not in school. The 'in_school' variable captures attendance at government, private, madarsa, or other types of schools, while the 'not_in_school' variable identifies children who have never enrolled or have dropped out. Using these, I defined my final sample selection variable, 'school_attendance_sample', which includes only those children currently attending school and not simultaneously categorized as out of school.

I generated dummy variables to distinguish between primary and secondary school students. This categorization was crucial for understanding how the effects of various factors might differ across educational levels. I created 'in_primary_school' for students in classes 1 to 5, and 'in_secondary_school' for those in classes 6 to 12.

Recognizing the importance of distinguishing between various educational institutions, I generated a set of dummy variables to represent government schools, private schools, madarsas, and other educational establishments. This approach allowed for a more granular examination of the characteristics and outcomes associated with each school category.

To capture regional variations, I created dummy variables for eleven key states that were particularly relevant to my research questions. These states were Kerala, Tamil Nadu, Bihar, Uttar Pradesh, Maharashtra, Rajasthan, Madhya Pradesh, Gujarat, West Bengal, Himachal Pradesh, and Assam. I assigned each of these states a unique numerical identifier and grouped the remaining states into a separate category. This allowed me to control for state-specific effects in my analyses.

My primary estimation strategy employed Ordinary Least Squares (OLS) regression with extensive controls for demographic and socioeconomic factors. To facilitate consistent inclusion of control variables across different model specifications, I used global macros in Stata to define a standard set of controls.

To address potential endogeneity concerns, I implemented family fixed effects models. These models allowed me to control for unobserved family-level characteristics that might influence both the decision to enroll a child in tuition and their academic outcomes. I used the 'xtset' command in Stata to set the household ID as the panel variable, enabling the fixed effects estimation.

I also employed ordered logit models to account for the ordinal nature of the outcome variables (reading and math proficiency scores). Following these models, I conducted marginal effects analysis to provide more interpretable results, particularly for policy implications.

To ensure the robustness of my results, I considered alternative model specifications. I ran separate analyses for different school types (government and private) and educational levels (primary and secondary). This stratified approach provided insights into how the effectiveness of tuition might differ based on the institutional setting and stage of education.

I also explored heterogeneity in tutoring effects by including interaction terms between the tuition dummy variable and other key variables such as gender. This allowed me to examine whether the impact of tutoring varies across different subgroups of students.

In addition to examining the binary effect of receiving tuition, I investigated the impact of the amount spent on tuition. For this analysis, I used the logarithm of tuition amount to account for the potentially non-linear relationship between tuition spending and academic performance, and to reduce the influence of extreme values.

While instrumental variable approaches can be valuable in addressing endogeneity, I did not employ this method in my primary analysis due to the lack of a suitable instrument in the ASER data. However, it's worth noting that alternative approaches have been used in similar contexts. For instance, Sharma (2009) used "the number of private schools available in the child's area of residence" as an instrument for private school attendance in a study on educational achievement in Nepal. While such an approach assumes that the number of private schools in an area is plausibly exogenous to an individual family's school choice, it may reflect the collective choices of parents in the area. Future research on shadow education in India could explore similar instrumental variable strategies if appropriate data becomes available.

The ASER data I used did not contain any personally identifiable information about the surveyed children or households. All analysis was conducted on anonymized data, with individuals identified only by randomly assigned household and child IDs. Throughout the analysis, I ensured that no individual-level data was reported or could be inferred from the results.

To ensure replicability, I have provided detailed information about the data source, variable definitions, sample selection criteria, and estimation methods in my thesis. The Stata code used for the analysis is well-commented and organized, allowing other researchers to replicate my results or extend the analysis.