



Predicting academic success: machine learning analysis of student, parental, and school efforts

Xin Jin¹

Received: 21 March 2022 / Revised: 27 October 2023 / Accepted: 29 October 2023
© The Author(s) 2023

Abstract

Understanding what predicts students' educational outcomes is crucial to promoting quality education and implementing effective policies. This study proposes that the efforts of students, parents, and schools are interrelated and collectively contribute to determining academic achievements. Using data from the China Education Panel Survey conducted between 2013 and 2015, this study employs four widely used machine learning techniques, namely, Lasso, Random Forest, AdaBoost, and Support Vector Regression, which are effective for prediction tasks—to explore the predictive power of individual predictors and variable categories. The effort exerted by each group has varying impacts on academic exam results, with parents' demanding requirements being the most significant individual predictor of academic performance; the category of school effort has a greater impact than parental and student effort when controlling for various social-origin-based characteristics; and significant gender differences among junior high students in China, with school effort exhibiting a greater impact on academic achievement for girls than for boys, and parental effort showing a greater impact for boys than for girls. This study advances the understanding of the role of effort as an independent factor in the learning process, theoretically and empirically. The findings have substantial implications for education policies aimed at enhancing school effort, emphasizing the need for gender-specific interventions to improve academic performance for all students.

Keywords Academic achievement · Machine learning · School effort · Family involvement · Gender disparities

Introduction

The literature has established that the academic achievements of students are influenced by the efforts exerted by the agents involved in the education process, namely, the school attended, the parents lived with, and the students. Despite the importance of effort, research on educational achievement has not adequately investigated the role of effort as an independent input in the education process, theoretically and empirically. Although student effort (e.g., subjective perceived effort, objective time spent learning) plays a crucial role in educational outcomes, parental effort (e.g., family involvement, parents' interest in their children) and school effort (e.g., classroom instruction, school management) are also vital (Gamboa & Waltenberg, 2012; Edmark & Persson,

2021; Golley & Kong, 2018; Broer et al., 2019; Dietrich et al., 2021). However, not every effort is equally important in achieving the desired outcomes. Notably, the literature has not investigated each effort's relative importance in explaining academic achievements.

This study aims to answer two questions: what is the relative importance of the individual effort variables in explaining academic achievements, and because some fields of influence appear more critical than others in predicting the outcome variable (Adler et al., 2018), what is the collective importance of these fields of influence. These answers will demonstrate the relative contribution of the effort variable group, which comprises school effort, parental effort, and student effort, in explaining the academic achievements.

To answer these questions, I use the China Education Panel Survey 2013–2015, a rich dataset that follows a cohort of sampled students throughout their junior high school years. The data are collected through comprehensive questionnaires completed by students in 7th and 9th grades, providing detailed information on their attitudes toward school and education. To construct measures of effort

✉ Xin Jin
xinjin@zedat.fu-berlin.de

¹ Fachbereich Erziehungswissenschaft Und Psychologie, Freie Universität Berlin, Fabeckstraße 37 & 69, Habelschwerdter Allee 45, 14195 Berlin, Germany

exerted by the different agents in the education process, I use a range of indicators. For students, I use self-reported answers to questions such as their reported time spent studying and their motivation to set and work toward academic goals. For parents, I examine their level of involvement in their children's education, such as whether they supervise their child's homework or frequently attend meetings with teachers. To capture the level of effort exerted by schools in the education process, I adopt 15 constructed effort variables, such as the implementation of interventions that support academic growth and development by the school, the availability of academic guidance services for students, and the type of disciplinary methods employed by the school's administration.

However, the empirical challenge is selecting economically and statistically significant effort variables that predict academic achievements among an adequate number of potential predictors. Traditional statistical methods techniques such as least squares regression (OLS) are limited in achieving accurate variable selection and good out-of-sample performance, especially when the number of regressors is large (Steyerberg & Harrell, 2016). Furthermore, the complex interplay between effort variables and academic performance might not be captured by parametric assumptions (Roick & Ringeisen, 2017), necessitating the use of estimation techniques that offer enhanced flexibility.

Machine learning techniques are ideal for addressing these challenges and answering this paper's two research questions. These techniques select influential features and model complex relationships between input variables and outcomes (Takeda et al., 2013). They are also ideal for managing high-dimensional data with many predictors while avoiding overfitting, a common problem in traditional statistical methods. In this study, I leverage four state-of-the-art machine learning tools—Lasso, Random Forest, AdaBoost, and Support Vector Regression (SVR)—to predict which individual effort variable or effort group is most predictive of academic performance. Lasso is a widely used regularization regression method for variable selection, which helps identify the most relevant effort variables. The remaining three methods model nonlinear relationships, which is crucial because of the complex interplay among varying effort variables and academic performance. Additionally, these methods rank variables by their prediction power, providing intuitive comparisons of the relative importance of different effort variables. Overall, using machine learning techniques, I perform a comprehensive analysis of many potential predictors and identify the most relevant effort variables for predicting academic achievement.

To assess the relative importance of each effort variable, I employ the aforementioned machine learning algorithms, using all 45 effort variables and 15 controlled variables to predict academic performance. I identify the top 20 most

important effort variables based on the coefficient magnitudes in Lasso and SVR, SHapley Additive exPlanations (SHAP) value in Random Forest, and importance score in AdaBoost. Among the top 20 predictors, most variables relate to school effort and parental effort, and some variables relate to student effort. Parents' demanding requirement is the most significant predictor among all individual variables. Furthermore, students' and parents' educational expectations exert a greater influence on academic achievements than other factors do. The practice of inviting parents to school events is the third most predictive factor for students' grades. These results indicate that if parents and schools prioritize education and highly value academic achievement, they may be more likely to provide a supportive (both at home and school) environment and encourage students' academic pursuits (Gbollie & Keamu, 2017).

To assess the importance of effort-related group variables, I include school effort, parental effort, and student effort variables independently in the model to predict academic performance. The results indicate that "school effort" is the most influential predictor of academic achievement, followed by parental effort, and students' effort has a limited impact. These findings underscore the crucial role of a supportive school environment, namely, school events and teacher supervision, in promoting academic success (Park et al., 2017). Furthermore, the heterogeneity test between male and female students finds that for girls, school effort has a greater impact on academic achievement than parental effort does; the opposite is true for boys. Thus, particularly in China, where parental investments often favor boys in multi-children families (Ling, 2017), girls who receive more attention and financial incentives from schools and teachers can be more equipped to navigate these changes and achieve academic excellence (Tang & Horta, 2021). These findings suggest that gender-specific interventions and support programs are necessary to improve academic outcomes for girls during the critical period of intellectual and academic development of junior high school.

Methodologically, this study contributes to the understanding of the relative contribution of multiple variables to students' academic achievements by using modern machine learning models (Masci et al., 2018). Traditional statistical models may provide biased results due to the unknown functional form of how effort affects grades, potential interactions between effort variables, and collinearity. By contrast, machine learning techniques offer flexibility, feature selection, model validation, and robustness to multicollinearity (Ogutu et al., 2012). Thus, the machine learning approaches in this study obtain good out-of-sample prediction accuracy by selecting relevant variables and reducing overfitting (Dalalyan et al., 2017). This strategy provides a feasible and superior approach to narrowing outcome predictors, especially in the case of large high-dimensional databases

in educational research. The ability to accurately predict unequal educational outcomes deepens the understanding of the effort-related factors that drive educational success and clarifies a strategic direction for additional compensation and policy intervention.

The findings of this study contribute to the literature by providing empirical evidence for the theoretical prediction that differences in how parents, schools, and students perceive and act in achieving higher academic performance can lead to disparities in academic outcomes (e.g., Edossa et al., 2018). This study emphasizes that academic success is not solely determined by objective structural factors, such as family background and school resources (Berkowitz et al., 2017) but also by latent motivation and tangible action efforts (Gneezy et al., 2019). As such, this study underscores the importance of increasing the effort to improve academic results and highlights the necessity to stimulate effort as a more feasible and effective approach than modifying social background or school resource allocation (Yeager & Dweck, 2012).

The identification of school effort as the most significant predictor of academic success has critical policy implications. Policymakers can focus on promoting various forms of school effort, such as creating a supportive and positive learning environment, providing individualized tutoring for students, encouraging teacher supervision, and enhancing student and parental participation in school events. By prioritizing school effort, policymakers can more effectively improve academic achievement than by relying solely on material resources or higher-quality teaching faculty and facilities.

The paper is organized as follows. The theoretical backgrounds are illustrated in Sect. 2, and the data source and variables are briefly introduced in Sect. 3. The methodology includes benchmark methods, feature extraction principles, and machine learning techniques and is presented in Sect. 4. The relevant results and discussions are elaborated in Sect. 5. The conclusion and potential policy implications are provided in Sect. 6.

Theoretical background

Sociologists have long been concerned with the extent to which inequality of opportunity, caused by circumstantial factors and family endowment, contributes to inequality of outcomes. Blau and Duncan (1967) were the first to establish a dual-driven theoretical model of family resource investment and self-motivated effort from a micro perspective. They proposed the status attainment model, using path analysis to explore the extent to which the occupational

attainment of the population in the United States is influenced by their family background and level of education at the micro-level (Ganzeboom et al., 1991; Winship, 1992). They regarded an individual's academic status attainment as the result of multiple factors that emerge sequentially throughout their life cycle; thus, they developed a pathway model incorporating innate and self-induced elements and inter- and intra-generational mobility into the analysis.

Although the classical status attainment model has been developed from structural and psychosocial perspectives, this study argues that several concerns are still worth discussing and expanding. Drawing on Bourdieu's (1984) theory of habitus and cultural capital, and Baumrind's (1971), Lareau's (2002), and others' studies of family parenting styles and school effectiveness, effortful devotions, namely, cognitive capacity, non-cognitive motivation, and observable time-devoted, may also be considered unavoidable factors impacting on educational outcomes (Deluca & Rosenbaum, 2001; Guan et al., 2006; Inzlicht et al., 2018; Shenhav et al., 2021). However, status attainment research has mainly disregarded the effort factor, which treats human capital invested in education as an effort factor concerning the family background to explain offspring educational outcomes (Caldas & Bankston, 1997; Sewell & Shah, 1968; Sheldon & Epstein, 2005). Human capital input is not equivalent to the individual effort factor; it primarily serves as a transmission and mediator between paternal and offspring status (Bourdieu, 2002; Kohn et al., 1990). The effects of actual psychological and behavioral efforts as independent exogenous variables and the mechanisms via which they function have not been examined.

To improve the understanding of the actual psychological and behavioral efforts exerted during task performance, Steele's (2020) framework on effort, which distinguishes between objective and subjective effort, offers valuable insights. Steele (2020) defined "objective effort" as tangible and measurable actions that reflect the amount of energy or work invested in a task, and "subjective effort" encompasses intangible internal experiences and attitudes associated with a task or goal. To operationalize effort in the context of this study, I adopt Steele's (2020) definition of effort. In this study, effort is examined at the individual student level, and at the level of parents and schools. Specifically, students' effort can be observed in how they approach education broadly, how they respond to classroom interactions with teachers, how much time they dedicate to and how much motivation they have for learning, how much inspiration they receive from family, how they conduct the necessary tests, and some other objective and perceived effort exerted to meet academic needs (Dunlosky et al., 2020; Mudrak et al., 2021). Similarly, by viewing "parents" and "schools" as behavioral agents in the same manner

as individual students, their efforts to improve students' educational outcomes during the task can also be defined as objective and subjective efforts rather than solely focusing on educational investment behaviors. These efforts can include psychologically devoting attention, stimulating motivation, instilling a sense of belief, and behaviorally spending additional time and energy on academic tasks (Stables et al., 2014; Ng & Wei, 2020).

These various efforts, shaped by family or school, modify educational behaviors that result in varying levels of academic performance, and increases social status (Burić & Sorić, 2012; Zimmerman, 2013). Efforts and effort-based capability can also supplement outcome disparities caused by structural factors when acting in different directions and with different forces (Darling-Hammond, 2018). If students inherently believe in devoting attention, parents and relevant schools would spend more time and energy on academic tasks, and the student's favorable outcomes would increase. Enhancing students' educational success is challenging, if not inconceivable, if the three key agents do exert the effort, regardless of the student's family background or school quality (Richardson et al., 2012). By taking a more nuanced approach than that in the literature to the role of effort in educational outcomes, understanding how different factors interact to influence student achievement can improve.

Therefore, highlighting the potentiality and capability of efforts to reduce outcome inequalities is rational. The learning motivations and exertions of students, parents, and schools can, to some extent, complement, compensate, and counter structural disadvantages in achieving equal outcomes (Amis et al., 2020; Hirsch, 2019). Additionally, students' acquisition of social and academic status is assuredly an integrated process affected by circumstantial and effort-related factors (e.g., Hodge et al., 2018) and a final collaborative result among efforts of parents, schools, and individuals (De Fraja et al., 2010). The distinction among the three agents' efforts is more akin to a spectral range than a dividing line; in reality, every action can be determined by a combination of these three components. An overemphasis on the influence of one level of factors at the expense of others may lead to reductionism or ecological fallacy in methodology (Curran & Bauer, 2011). Therefore, a thorough analysis of the three sides must be considered, especially to examine which aspect dominates students' learning progress, resulting in disparities in student achievements under an integrated framework. More specifically, this study aims to determine the extent to which factors can have the most predictive effects on educational outcomes, while all three types of efforts are considered simultaneously. The conceptual framework for this study is illustrated in Fig. 1, which provides a comprehensive overview of the research progress.

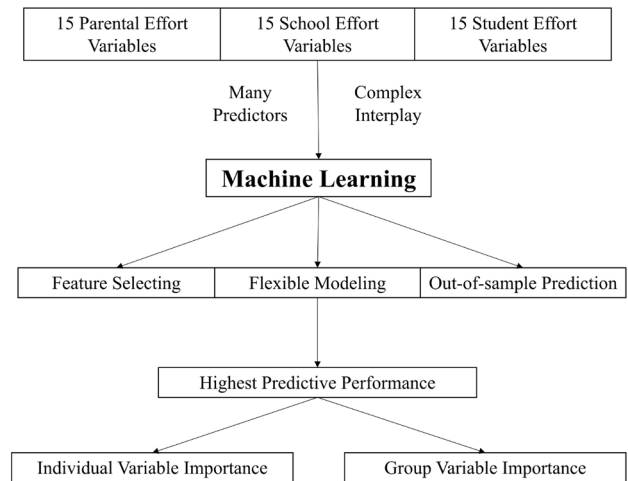


Fig. 1 Conceptualized framework

Data and measures

Data source

This study uses data from the China Education Panel Survey (CEPS), a nationally representative survey designed and implemented by Renmin University in China. The primary aim of the survey is to investigate the impact of various factors, namely, family, school, individual, and macro-social structures, on students' academic achievements. The survey was conducted by selecting a sample of 112 schools, 438 classes, and approximately 30,000 students by using a national sampling method.

The large sample size of the CEPS is a substantial strength of this study because it generates more accurate averages, identifies outliers, and yields reduced margins of error (Wang, 2016), enhancing the external validity of the findings. Moreover, the survey provides detailed information on three key agents' efforts and demographic characteristics, namely, individual innate ability, family background, and school resources, which are essential for understanding the internal and external environments of students (Ma & Wu, 2019; Xu, 2016) and, thus, this analysis.

I drop missing values for all relevant variables and remove extreme values to avoid potential bias from outliers. The final estimation sample comprised 24,974 students' information.

Measures

Dependent variable

Academic achievement To measure academic achievements, I use students' total test scores: the sum of Chinese, Mathematics, and English scores. The data were sourced

Table 1 Descriptive statistics

Variables	Abbr	Mean	SD	Min	Max
Academic achievement (total scores)	TotalScore	236.4	74.38	0	440
Student effort					
Time spent studying and completing homework assignments					
Time spent completing in-class homework	StuSchoolHomework	5.617	3.614	0	48
Time spent completing ex-class homework	StuExtraHomework	2.244	2.688	0	48
Time spent attending cram school	StuCramSchool	1.969	2.659	0	48
Self-perceived subjective effort					
Student self-dedication	StuDedication	3.217	0.847	1	4
Student self-persistence	StuPersistence	3.260	0.843	1	4
Student self-resilience	StuResilience	3.273	0.890	1	4
Seeking out additional help or resources when needed					
Attend tuition classes (related to schoolwork) (no=0, yes=1)	StuTuition	0.318	0.466	0	1
Participate in summer/winter camps (no=0, yes=1)	StuSummerCamp	0.103	0.304	0	1
Setting and working toward academic goals					
Student self-expectation	StuExpectation	6.891	1.761	1	10
Student self-confidence	StuFaith	3.176	0.721	1	4
Engaging in extracurricular activities that support academic growth and development					
Attend International Mathematical Olympiad (IMO) class	StuOlympiad	0.036	0.187	0	1
Attend extra Mathematics (exclude IMO) class	StuExtraMath	0.214	0.410	0	1
Attend extra Chinese class	StuExtraChinese	0.109	0.311	0	1
Attend extra English class	StuExtraEnglish	0.231	0.421	0	1
Frequency of visits to museums	StuVisitFreq	1.990	1.123	1	6
Parental effort					
Engaging in their child's academic growth and development					
Help with their child's homework	ParTutoring	2.192	1.142	1	5
Supervise their child's homework	ParMonitor	1.649	1.123	0	4
Frequency of parental visits to museums with their Child's	ParVisitFreq	2.085	1.220	1	6
Communicating regularly with teachers and staying involved in their child's academic progress					
Parents talk to teachers about their child's learning	ParCareLearning	0.691	0.462	0	1
Whether parents proactively contact with teachers	ParContactTeacher	2.373	1.020	1	4
Parents' attitudes toward their child's academic performance					
Parental discipline for their child	ParDiscipline	0.659	0.474	0	1
Parents enrolling their child in tuition	ParTuition	0.092	0.290	0	1
Parents' concern for their child's effort level	ParPerception	0.857	0.350	0	1
Parents' academic goals and career aspirations for their Child's					
Parents' requirements for their child's performance	ParRequirement	3.018	0.858	1	4
Parents' educational expectations for their child	ParExpectation	6.957	1.577	1	9
Parents' faith in their child	ParFaith	3.227	0.689	1	4
Modeling good study habits and time management skills at home					
Parents being strict about their child's homework and exams	ParCareExams	2.363	0.532	1	3
Parents being strict about their child's school behaviors	ParCareBehavior	2.283	0.586	1	3
Parents being strict about their child's time spent on the internet	ParBanInternet	2.581	0.566	1	3
Parents being strict about their child's time spent watching TV	ParBanTV	2.361	0.586	1	3
School effort					
Implementing supportive interventions that support academic growth and development					
School requires students to attend night study	SchNightStudy	1.563	0.715	1	3
Teachers on duty for night study	SchSupervision	0.951	0.216	0	1
School organizes summer/winter camps for students	SchSummerCamp	0.103	0.304	0	1
Supporting teacher professional development and addressing student academic/life needs					
Frequency of school sessions on academic/life coaching	SchCoaching	2.251	0.842	1	4

Table 1 (continued)

Variables	Abbr	Mean	SD	Min	Max
Availability of teacher training	SchTeacherTraining	1	0	1	1
Partnerships with local businesses for additional resources	SchPartnership	0.086	0.281	0	1
Encouraging parent and community involvement in school activities and events					
Frequency of parent-teacher meetings	SchParentMeeting	2.628	0.615	1	4
Frequency of written reports from the school to parents	SchWrittenReport	2.769	0.861	1	4
Frequency of schools inviting parents to observe	SchClassReport	2.067	0.893	1	4
Providing effective and engaging classroom instruction					
Main teaching methods: teacher-led lectures	SchTeacherLecture	0.925	0.264	0	1
Main teaching methods: group discussions	SchGroupDiscussion	0.586	0.493	0	1
Main teaching methods: bilingual teaching	SchBilingualTeaching	0.062	0.242	0	1
Main teaching methods: stratified teaching	SchStratifiedTeaching	0.090	0.285	0	1
Offering individualized academic support services such as tutoring or academic counseling					
School offers remedial classes for students with failing grades	SchRemedialCourse	1.981	1.367	0	4
School offers advanced study for students good at a single subject	SchImprovedCourse	0.572	0.495	0	1
<i>N</i>	24,974				

Descriptive statistics for the variables used in the pooled ordinary least squares (OLS) and machine learning regression analyses. For brevity, all abbreviations used in the table refer to the aforementioned regression results. Specifically, variable abbreviations with the prefix “Stu-” denotes student effort, “Par-” denotes parental effort, and “Sch-” denotes school efforts. Detailed descriptive statistics of the controlled variables used in the analysis are given in Table 4

from the students’ term exam scores across two consecutive school years and provided by their respective schools. Table 1 shows that the average score for these students is 236 points, accounting for 52.4% of the maximum possible score of 450 points.

Predictors

Student effort Psychological and behavioral efforts play a crucial role in improving educational attainment (Schunk & DiBenedetto, 2020). Therefore, the effort students invest in their academic work is a significant factor influencing their academic achievement. To assess student effort, I use multiple survey measures, namely, students’ (1) self-reported time spent studying and completing homework assignments, (2) subjective perception of their effort levels, (3) proactive efforts to seek additional help or resources when needed, (4) motivation to set and pursue academic goals, and (5) engagement in extracurricular activities that promotes academic growth and development. These measures are encoded into categorical variables, with higher values representing a greater level of student effort.¹ I include 15 proxies for student effort.

Parental effort Parental effort is assessed based on parents’ level of involvement and support in their child’s education, and their attitudes toward their child’s academic performance (Avvisati* et al., 2010). Specifically, this study measures parental effort by using four variables: parents’ (1)

engagement in their child’s studies, (2) willingness to discuss their child’s progress with teachers, (3) academic goals and career aspirations for their child, and (4) role in modeling good study habits and time management skills at home. Higher values on these measures indicate a higher level of parental effort in contributing to their child’s educational success. I include 15 variables to measure parental effort.

School effort This study measures school effort by using five indicators related to activities that extend beyond the mandatory requirements of educational institutions (Baños et al., 2019): (1) implementation of interventions that support academic growth and development; (2) parent and community involvement in school activities and events; (3) provision of academic and life guidance to students; (4) practice of grouping students based on similar abilities; and (5) disciplinary methods employed by schools, such as offering night study sessions or individualized academic tutoring. Higher values on these measures indicate greater school effort in fostering students’ academic success. I include 15 school effort variables.

Descriptive statistics

Methodology

This study incorporates 45 effort variables as the key independent variables. Understanding the relative contribution of each variable to students’ academic performance is empirically challenging. First, the functional form of how effort

¹ Details on how these variables are constructed are in Table 3 in the Appendices.

affects grades is unknown. Various efforts may interact and have nonlinear effects on a student’s academic performance. Assuming a simple, additive linear model using conventional OLS imposes strong parametric assumptions and might provide biased results. Multiple variables may exhibit collinearity, making isolating their marginal effects difficult. To alleviate these concerns, I employ machine learning techniques, which offer several benefits over traditional statistical approaches.

- (1) Flexibility: Machine learning algorithms can learn complex and nonlinear relationships between independent and dependent variables without imposing strict assumptions.
- (2) Feature selection: Machine learning can automatically identify the most relevant variables among the 45 effort variables, providing a more concise and interpretable model than those in the literature.
- (3) Model validation: Machine learning models leverage techniques such as cross-validation, which helps ensure the external validation of the findings and reduces the risk of overfitting.
- (4) Robustness to multicollinearity: Machine learning methods, such as regularization techniques, can manage situations where predictor variables exhibit collinearity, mitigating the adverse effects on the model’s performance.

By leveraging these advantages, machine learning techniques enable a more nuanced exploration of the relationship between various effort variables and students’ academic achievements, ultimately deepening the understanding of the factors that drive educational success.

Benchmark model: OLS

To investigate the relationship between effort factors and students’ academic achievement, I first estimate the following baseline linear regression model:

$$y_i = StudentEffort\beta_1 + ParentalEffort\beta_2 + SchoolEffort\beta_3 + X\sigma + u_i, \# \tag{1}$$

where y_i is the total test scores of student i , and **StudentEffort**, **ParentalEffort**, and **SchoolEffort** are vectors that include all student effort, parental effort, and school effort variables, respectively. X is a control variable vector, namely, students’ demographics, parents’ background characteristics, class-fixed effects, and year-fixed effects. By integrating control variables in the regression, the comparison can be restricted to students with similar characteristics, which improves the precision of estimates of the effect of effort factors. u_i is the error term. The regression equation was estimated using ordinary least squares (OLS). To make

the coefficients comparable, I normalize all the right-hand-side variables. Thus, the coefficients of the effort variables can be interpreted as the change in academic achievements associated with a one standard deviation change in the corresponding effort variable. I use this normalization procedure to compare the effects of different types of effort variables on academic performance in a standardized manner.

Individual variable importance using machine learning tools

In this section, I use multiple machine learning techniques to examine the explanatory power of each effort variable. I first provide a brief introduction to the machine learning models used and then explain the analysis procedure.

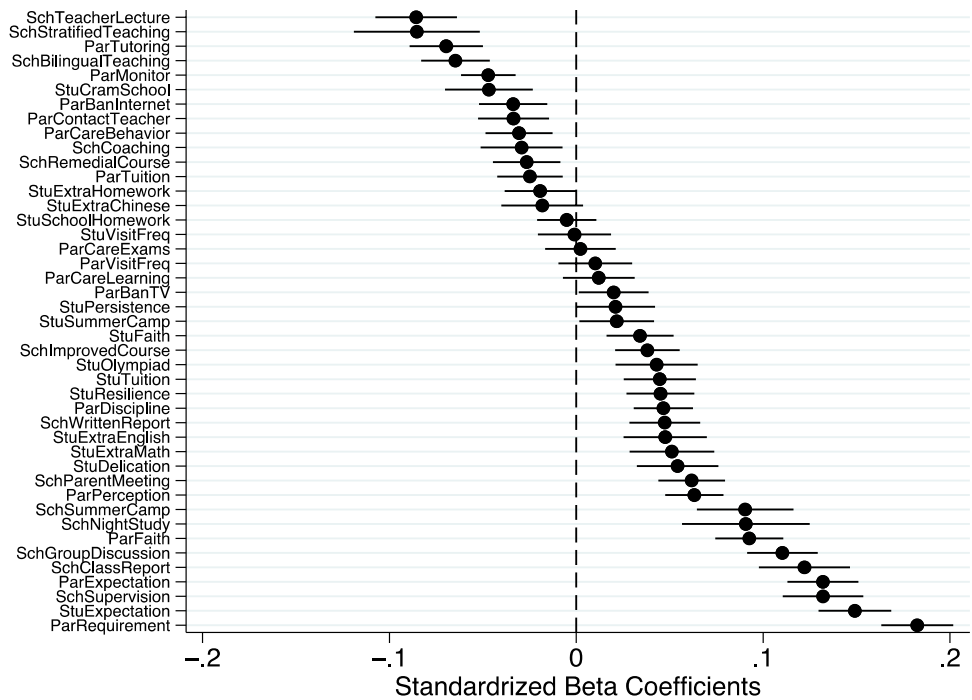
Lasso The first method used is Lasso, a widely used regularization regression technique. Lasso regression performs both feature selection and regularization to enhance the predictive accuracy and interpretability of statistical models. The objective function of Lasso is to minimize the following:

$$SSR + \lambda \sum_{j=1}^p |\beta_j| \# \tag{2}$$

Lasso is similar to regression in that it still requires the imposition of parametric assumptions. The first term that I minimize is the sum of squared residual (SSR), equivalent to regression. However, Lasso includes a penalty term (the second term) that ensures that it does not overfit the data and delivers good predictive performance under approximate sparsity. A key aspect of operationalizing Lasso is tuning the “complexity cost” λ , which involves selecting the appropriate value for the penalty level. The best practice is to use cross-validation to identify the optimal value for this hyperparameter.

Random forest The second method used is Random Forest, a tree-boosting method that achieves high prediction accuracy in many prediction tasks. Random forest is a flexible nonparametric model that can manage complex interactions among variables and is well suited for high-dimensional data. It works by building an ensemble of decision trees on random subsets of the data and variables. This approach helps reduce overfitting and improve the accuracy and robustness of the model. The final prediction is then made by averaging the predictions of all the decision trees in the ensemble. Random Forest also provides information on variable importance, which can help identify the most important predictors of academic outcomes. To avoid overfitting, Random Forest also has hyperparameters, such as the number of trees in the ensemble, the maximum depth of the trees, and the minimum number of samples required to split a node. Cross-validation is used to select the optimal values of these hyperparameters.

Fig. 2 Standardized beta coefficient plot of OLS estimated effects on academic achievements



AdaBoost The third method used is another ensemble method, AdaBoost, a boosting algorithm that iteratively combines weak classifiers to create a strong classifier. AdaBoost is effective in a wide range of prediction tasks and is particularly useful for identifying important predictors. It works by assigning higher weights to observations misclassified by the current set of weak classifiers, emphasizing these observations in the next round of classification. By iteratively improving the classification accuracy of the weak classifiers, AdaBoost creates a strong classifier that accurately predicts the outcome variable. One advantage of AdaBoost is its ability to identify important predictors by assigning higher weights to more informative variables for classification. This allows a focus on the most important variables and reduces the dimensionality of the data, which can improve the accuracy and interpretability of the model.

Support vector regression (SVR) The last method used is SVR. SVR constructs a hyperplane in a high-dimensional space that maximally separates the data points into two classes: one for the outcome variable below a certain threshold and the other for the outcome variable above the threshold. SVR is particularly useful for identifying important variables. By selecting the most informative variables for inclusion in the kernel function, which is used to transform the input variables into a higher-dimensional space, SVR can improve the predictive accuracy of the model while reducing the dimensionality of the data. Another advantage of SVR is its ability to manage nonlinear relationships between the input and the outcome variable. SVR achieves this improvement by using a kernel function where nonlinear

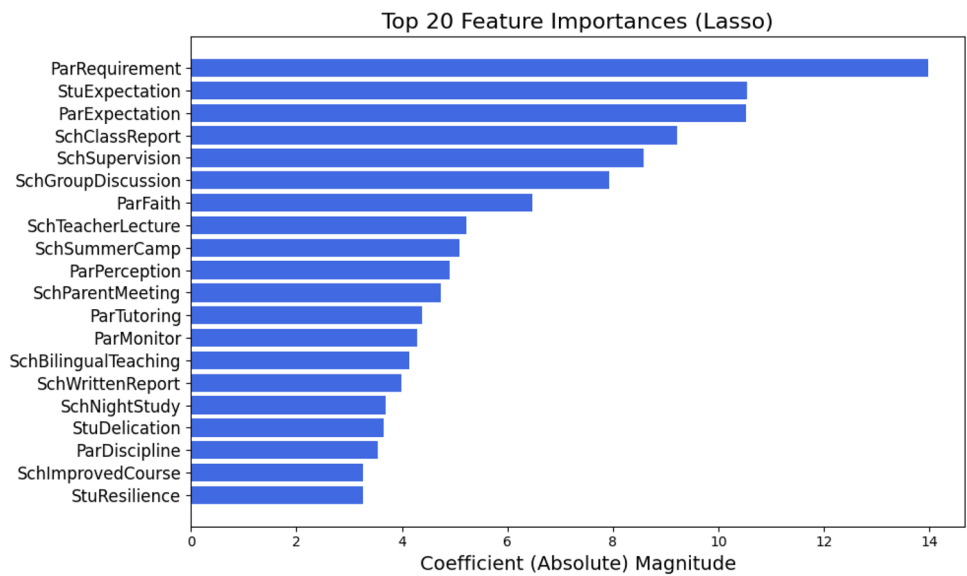
relationships can be more easily captured. Thus, SVR is a powerful tool for predicting continuous outcome variable and identifying the most important predictors.

Procedures for selecting the most important variables

To gain insights into the importance of individual variables, I use the following procedures:

- (1) All variables were standardized before analysis to ensure comparability.
- (2) The dataset was split into a training set (80%) and a test set (20%) to evaluate the performance of the models.
- (3) The aforementioned four machine learning models were trained on the training set, and hyperparameters were selected using tenfold cross-validation. In AdaBoost, decision trees were used as weak classifiers.
- (4) The feature importance was sorted in descending order, and the top 20 features were selected, excluding control variables. For Lasso and SVR, the absolute value of the coefficient magnitude was used to measure variable importance. For Random Forest, the mean absolute SHAP value was used. In AdaBoost, an importance score, calculated by summing the weights of the samples misclassified by the weak classifiers in each iteration of the boosting process, was used to measure variable importance.

Fig. 3 Coefficient (Absolute) magnitude of variables assessed by lasso



- (5) The test mean squared error (MSE) was computed to assess the goodness of fit of the models.

Group variable importance using machine learning tools

Procedures for assessing group variable importance

To investigate the relative importance of each variable group (i.e., student effort, parental effort, and school effort) in predicting academic outcomes, I use the following procedures:

- (1) Again, all variables were standardized, and the dataset was split into a training set (80%) and a test set (20%).
- (2) Machine learning models were trained using only the variables in each of the three groups separately: student effort, parental effort, and school effort. This method allowed for a direct comparison of the relative importance of each variable group in predicting academic outcomes.
- (3) The test MSE was computed for each separate model, with the variable group with a smaller test MSE indicating a higher model fit and greater importance of the variables in that group.

By comparing the test MSE across the models, I gained insights into the relative importance of each variable group in predicting academic outcomes. These results are suitable to inform educational policies and interventions aimed at improving academic performance, such as focusing on increasing parental involvement or improving teaching practices in schools.

Results and discussion

Benchmark model: OLS

Figure 2 presents the baseline OLS point estimates and 95% confidence intervals. To conserve space and avoid distraction from the focus of this analysis, I do not report the coefficients of control variables. Figure 2 suggests that several factors have a significantly positive impact on academic performance. Specifically, parents’ expectations for their child’s academic performance, student and parental expectations, and the presence of teachers during night study sessions have a positive and statistically significant effect on academic outcomes. Notably, parents’ expectations have the strongest positive influence on educational achievement.

In contrast, certain teaching methods, such as teacher-led lectures, stratified teaching, and bilingual teaching, have a statistically significant and negative effect on academic performance. Parents’ involvement in tutoring and supervision also has an adverse impact. Furthermore, the frequency of student and parent visits to museums, and parents’ strictness regarding homework and exams have negligible effects on academic achievement; their coefficients are centered around zero. Similarly, variables such as extra homework, attending extra Chinese classes or summer/winter camps, and self-perceived persistence and faith in learning, have minimal impact on academic performance; their coefficients are small.

However, interpreting the results with caution is essential. First, including too many independent variables in an OLS model can lead to overfitting and may result in non-significant predictors included in the model. Second, the

Fig. 4 Mean absolute SHAP value of variables assessed by random forest

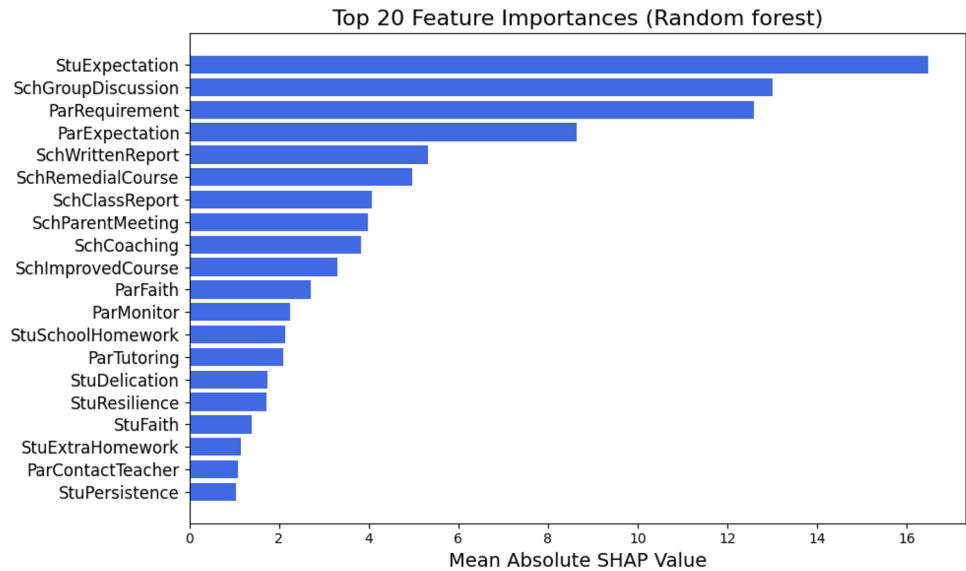
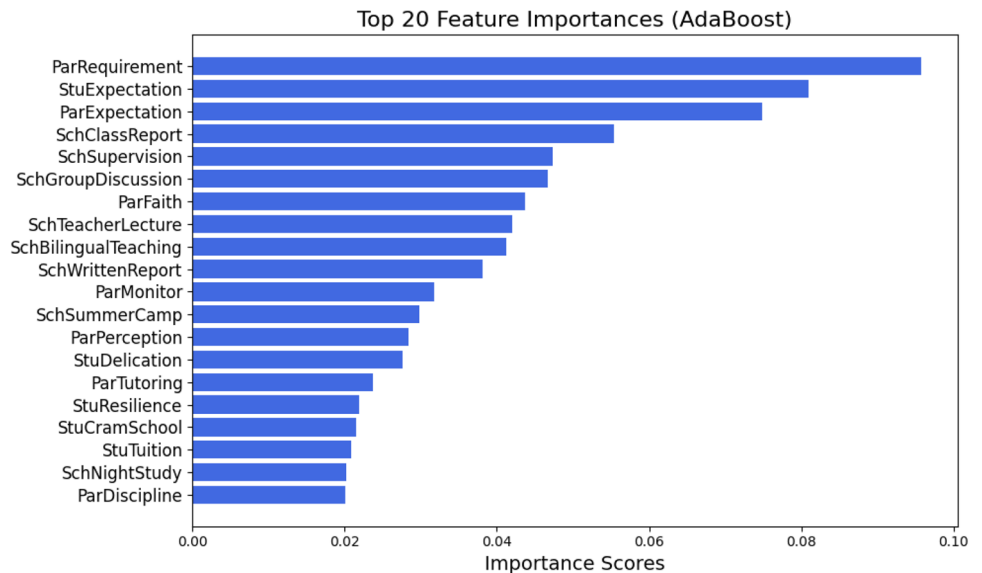


Fig. 5 Importance scores of variables assessed by ada boost



baseline linear model imposes strong parametric assumptions that may not hold in practice. Therefore, considering alternative methods that improve the capturing of the complexity of the data and identify the most important predictors of academic outcomes is crucial.

Individual variable importance

Figures 3, 4, 5, 6 display the top 20 important predictors for academic achievement as determined by four models: Lasso, Random Forest, AdaBoost, and SVR.

Lasso, AdaBoost, and SVR produce remarkably similar results. Figures 3, 5, and 6 reveal that the most important predictor is parents' requirements for their child's

academic performance (*ParRequirement*). Parents who set demanding requirements and actively engage in their child's education can provide valuable support and resources that contribute to their child's success (Boonk et al., 2018). Following *ParRequirement*, all three models predict that both students' expectations (*StuExpectation*) and parents' expectations (*ParExpectation*) are highly instrumental in forecasting academic achievement. This finding implies that students who possess higher levels of intrinsic motivation and receive encouragement from their parents tend to perform better academically (Ryan & Deci, 2020). Furthermore, the robust predictive power of schools' practice of inviting parents to attend school events (*SchClassReport*) and having teachers supervise

night study (*SchSupervision*) underscores the positive impact of a supportive school environment on students' academic success (Deming et al., 2014). Overall, these findings stress the importance of parental involvement and supportive school environments in promoting students' academic success. By prioritizing education and providing a supportive and engaging learning environment, parents and schools can help students reach their full potential (Gbollie & Keamu, 2017).

The Random Forest model produces similar predictors, although the relative rank differs slightly from that of the other three models. Figure 4 shows that student self-expectations (*StuExpectation*) rank first for feature importance, with a mean absolute SHAP value of more than 16. The school employing group discussion (*SchGroupDiscussion*) as a main teaching approach ranks second in determining students' academic performance because it can promote a collaborative learning environment that promotes critical thinking, communication, and problem-solving skills, leading to a more engaged and active learning experience (Al-Samarraie & Saeed, 2018). Parents' requirements (*ParRequirement*) and expectations (*ParExpectation*) in their child's academic records are also crucial, as shown in the previous three models.

An alternative method to interpret the results is examining the number of parental, school, and student variables ranked among the top 20 most important predictors. The machine learning models indicate that school effort is the most important factor, and student effort is the least important. For instance, in Lasso, among the top 20 predictors, 10 variables related to school effort, 7 to parental effort, and 3 to student effort. In SVR and AdaBoost, 8 variables pertained to school effort, with a higher relative rank among the top 20 predictors, and 5 variables related to student effort. Overall, school effort-related variables

Table 2 Comparison of machine learning model performance in predicting outcome variable

	Train MSE	Test MSE
Random forest	317	2318
Lasso	3176	3363
Support vector regression	3209	3377
AdaBoost	2914	3195

have greater predictive power. They are critical because they reflect the quality and effectiveness of the educational environment. A school that provides pupils with support and resources while promoting a positive and involved learning community is more likely to improve academic achievement than a school that does not focus on these traits (Berkowitz et al., 2017).

Although the analysis examines variable importance, comparing the four models based on their in-sample and out-of-sample performance is also important. Table 2 presents the MSE values on the training and test data, with a lower MSE value indicating better performance in predicting the outcome variable. The results demonstrate that the Random Forest model outperforms the other models in test MSE, with a relatively low value of 2318, suggesting that it fits the test data better than the other three models do. AdaBoost, another ensemble method, performs worse than Random Forest with a test MSE of 3195, although it performs slightly better than Lasso (test MSE = 3363) and SVR (test MSE = 3377). Because of its flexibility and strong performance, I use the Random Forest model to assess the importance of the group variable in the next section. I use Lasso regression as a robustness check because of its interpretability.

Fig. 6 Coefficient (Absolute) magnitude of variables assessed by SVR

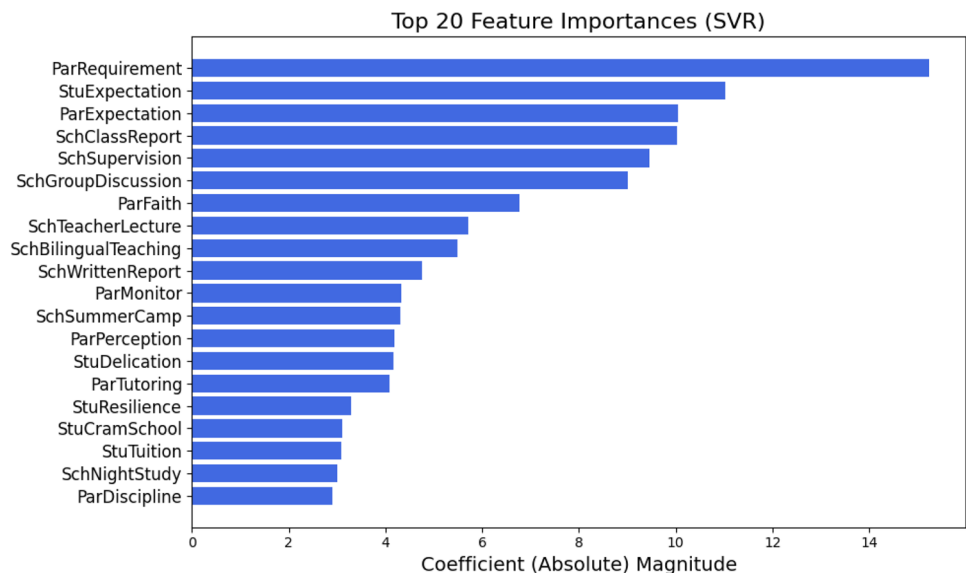


Fig. 7 Group variable importance assessed by random forest

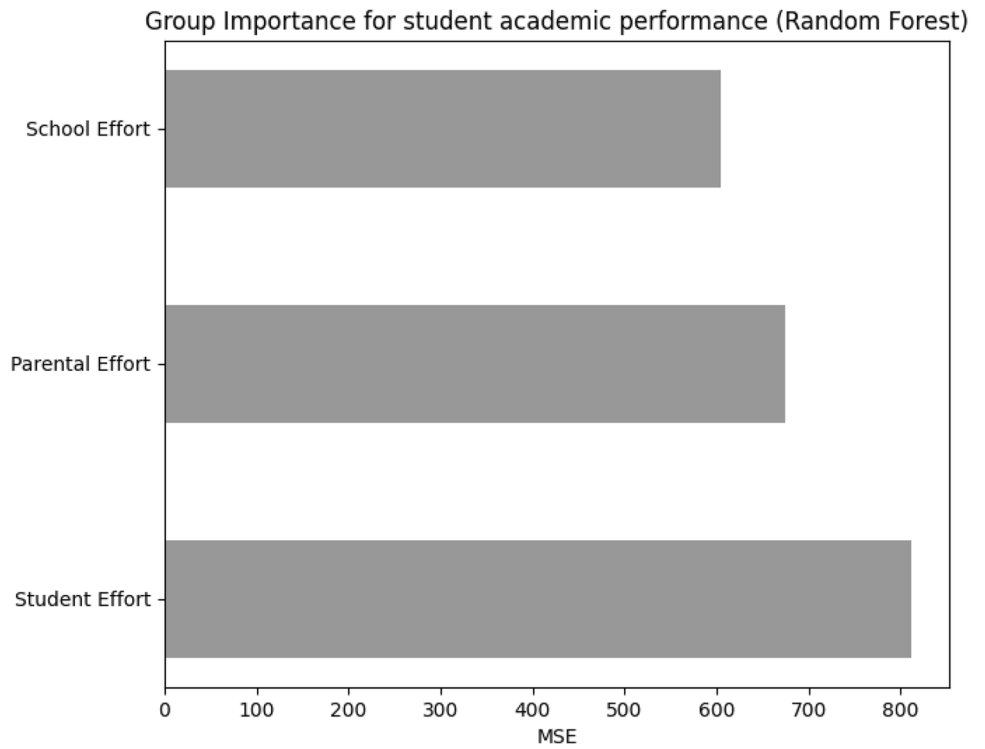


Fig. 8 Group variable importance assessed by lasso

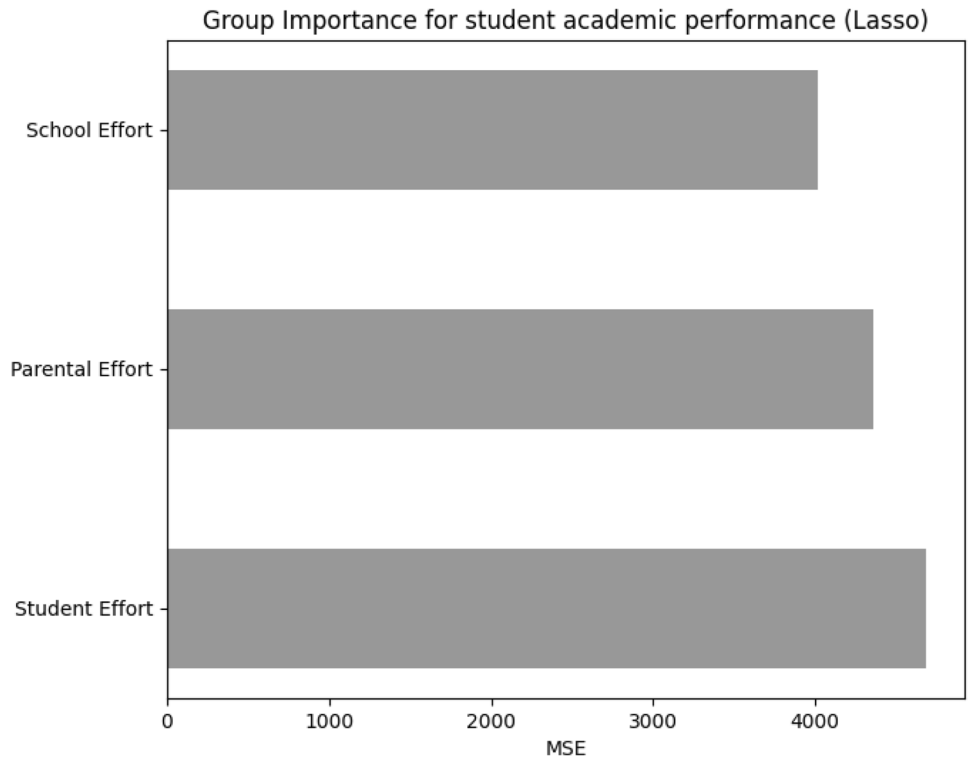
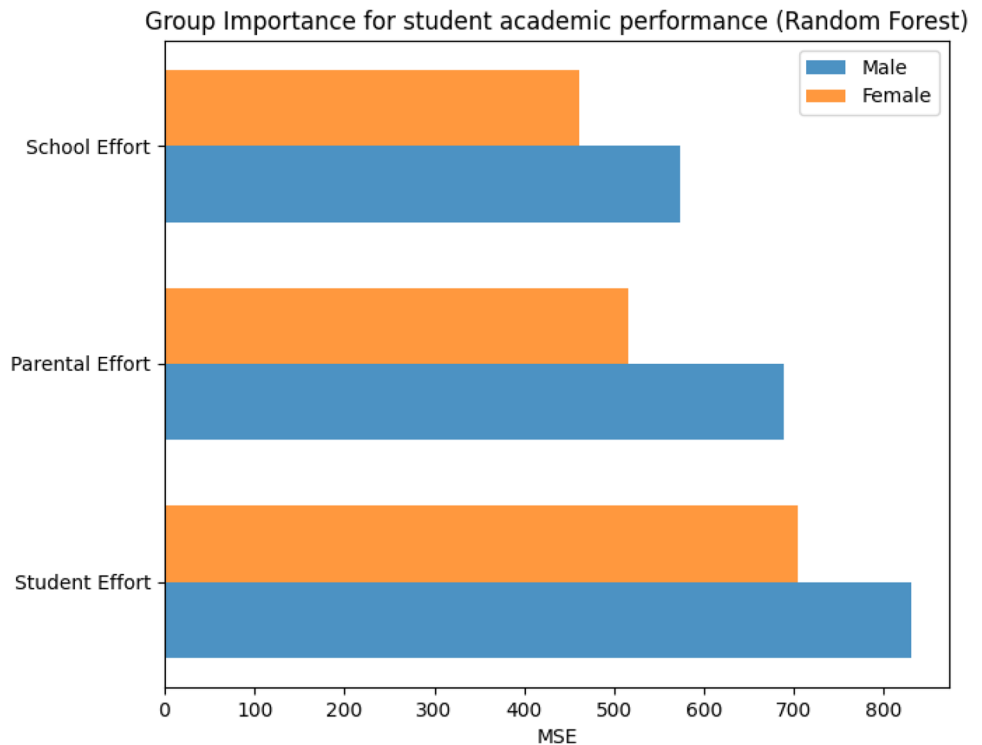


Fig. 9 Group variable importance assessed by random forest (by Gender)



Group variable importance

Figures 7 and 8 display the results of the variable group importance analysis. The Random Forest and Lasso models predict that the “school effort” category is the strongest predictor of educational outcomes, followed by parental effort and individual effort. Using school effort variables alone yields better model prediction (lower MSE) than using parental or student effort variables.

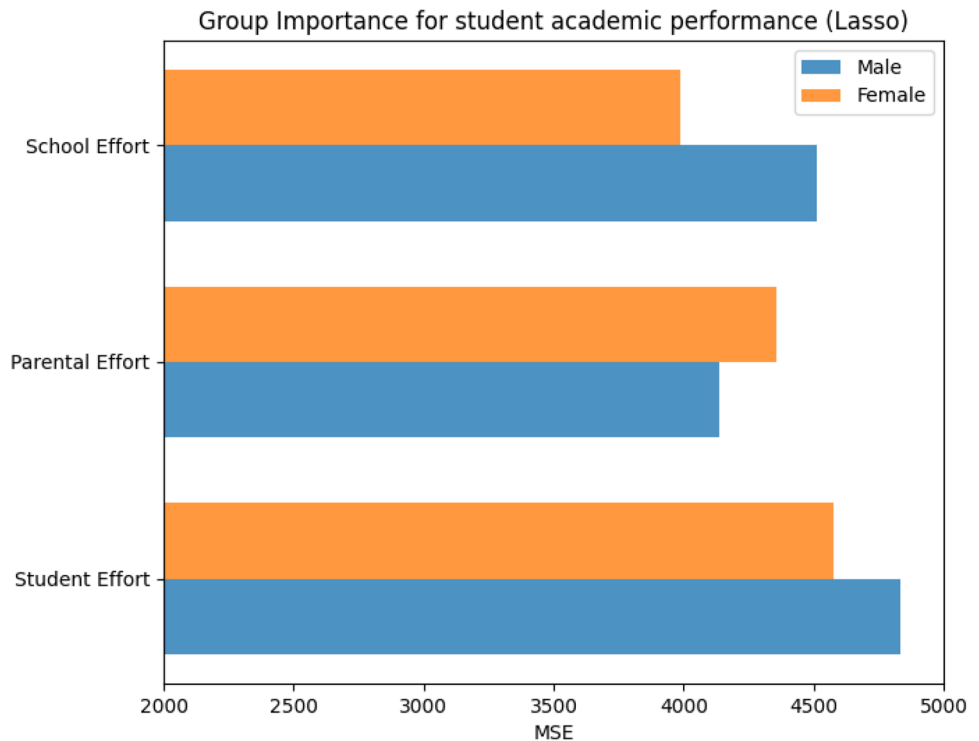
Schools that devote considerable effort are more likely to motivate and engage students in their academic work than those that do not, leading to better academic results. One possible explanation for this phenomenon is that a supportive school environment can create a sense of belonging and motivation among students, resulting in increased engagement and effort in their academic pursuits (Won et al., 2018). When parents are involved in school events and teachers provide academic support outside regular class time, students may perceive that the school community values and supports their education (Đurišić & Bunijevac, 2017). This perception can lead to improved academic performance because students are more likely to take their studies seriously and

strive for success. Additionally, the extra academic support and sense of community provided by a supportive school environment can help students overcome challenges and obstacles that may impede their academic progress (Darling-Hammond & Cook-Harvey, 2018).

Parental effort might be another critical group variable that affects students’ academic achievements, responding to studies that parents’ involvement and high expectations are incentives for academic improvement (Fang et al., 2018). Additionally, the demanding requirements set by parents may increase students’ learning performance. This result also can be explained as follows: more effective interaction between students and parents, as a critical part of high educational investments, leads to an increase in attention and improvement in support from family (Boonk et al., 2018). In addition, when parents convey their knowledge, attitudes, and disciplines toward learning, the student’s correspondingly improved performance will evoke or “demand” additional home instruction in a virtuous cycle (Soni & Kumari, 2017).

To investigate whether there are gender differences in the impact of effort levels on academic achievement, I conduct

Fig. 10 Group variable importance assessed by lasso (by Gender)



the same analysis on samples of male students and female students. Figures 9 and 10 present the results. For girls, in academic success, school effort is more significant than parental and individual effort. For boys, parental effort is the most important factor. This gender-based difference may explain the recent trend of higher academic achievement among girls than among boys, particularly in China, where parental investments are more likely to favor boys than girls in multi-children families (Ling, 2017). Conversely, schools and teachers are more likely to provide equal incentives and resources to both genders, creating a more level playing field for girls to excel academically (Tang & Horta, 2021; Verge, 2021).

Overall, this study provides valuable insights into the factors that impact academic achievement among junior high school students in China. The findings highlight the importance of group-level factors, specifically school effort,

especially academic support, in predicting academic success. A supportive school environment that engages and motivates students, and a school community that values and supports education, can have a significant impact on academic outcomes. Moreover, the study reveals gender differences in the effects of effort levels on academic achievement, with school effort being the most substantial factor for girls and parental effort being the most substantial factor for boys. Thus, policymakers aiming to improve academic performance should focus on stimulating school efforts, which is a more feasible and practical goal than attempting to change the social context of families or the resources of school hierarchies. By emphasizing the importance of school effort and parental involvement, policymakers can create a more supportive and conducive learning environment for students, improving academic outcomes and opportunities for success.

Conclusion and implications

Predicting educational outcomes is crucial to policy implementation and social development. This study uses machine learning techniques to provide insights into how parental, school, and individual efforts might shape and aggravate educational inequalities. This study stresses the importance of effort as a direct predictor of student academic outcomes. It considers effort a vital driver of upward mobility in social and educational settings. If rewarding students' effortful behaviors, such as increasing their determination, perseverance, and patience regarding learning, efforts can compensate for unbalanced educational resources gained from family backgrounds to sustain or upgrade the existing academic status and future social status. The core finding of this research also indicates that efforts from both parents and schools, whether they are analyzed through their distinct variables or perceived as two group variables, are identified as decisive factors in improving educational outcomes. Therefore, future social and educational inequalities studies must consider the potential for various efforts where distinct effects exist across socioeconomically heterogeneous groups. What might be faster and more reliable than waiting for their economic circumstances to improve is to encourage effort in groups with comparatively disadvantaged social or academic status.

This study emphasizes the critical role of school efforts in improving educational outcomes. By contrast, in China, the government has primarily focused on policies aimed at strengthening school resource-based effectiveness, such as the "Quality Education" initiative that invests in teacher training and educational materials (Wang et al., 2019) and the "Double Reduction" policy that transfers academic responsibilities to families (Eryong & Li, 2022). Schools should also prioritize implementing intramural motivational strategies to enhance students' self-efficacy and susceptibility to incentives (Hong et al., 2017). This approach would address outcome and effort disparities by improving the quality of in-school education across all stages of schooling. Findings in Zhu (2019) and Fu (2020) support the effectiveness of this approach. Therefore, schools should carefully and thoughtfully design and

implement motivational strategies to create a positive, supportive learning environment for their students.

On the basis of this research, I propose the following strategies to improve school efforts and students' academic performance:

- (1) Increase the number of internal school programs, such as workshops, assemblies, and classroom discussions, emphasizing the link between school effort and student academic achievement.
- (2) Provide personalized instruction, extracurricular events, and one-on-one support from tutors or teachers for students in need.
- (3) Create a positive learning atmosphere that motivates students to participate in their studies and be responsible for their academic growth. This strategy can be accomplished by promoting an engaging and inclusive school context, providing opportunities for student leadership and collaboration, and recognizing and celebrating students' achievements.

Furthermore, recognizing the potential limitations of a prediction task is critical, for example, correlation versus causation; thus, non-causal estimates based on statistical relationships between effort-related factors and students' academic achievements may not reliably identify the variables' underlying causal impacts. As a result, the findings of this study should be interpreted with caution and supplemented with other research methods, such as randomized controlled trials, to demonstrate causal correlations. Another possible limitation is that predictive models only forecast outcomes within the range of the data used to train them, resulting in erroneous extrapolations. Thus, additional data, including that from other sources, must be collected to provide a comprehensive picture of the predicted effort variables.

Appendix

See Table 3.

Table 3 Operationalizations of independent, dependent, and controlled variables

Variables	Operationalizations
Academic achievement	Summed Chinese, mathematics, and English test scores
Student effort	
Time spent completing in-class homework	The total time (in hours) a student spends completing in-class homework per week
Time spent completing ex-class homework	The total time (in hours) a student spends completing ex-class homework per week
Time spent attending cram school	The total time (in hours) a student spends attending cram school per week
Student self-dedication	The level of dedication a student reports for their academic work—"I would try my best to finish even the homework I dislike"—on a scale from 1 (Strongly disagree) to 4 (Strongly agree), with 1 being low persistence and 4 being high dedication
Student self-persistence	The level of persistence a student reports for their academic work—"I would try my best to finish my homework, even if it would take me quite a long time."—on a scale from 1 (Strongly disagree) to 4 (Strongly agree), with 1 being low persistence and 4 being high persistence
Student self-resilience	The level of resilience a student reports for their academic work—"I would try my best to go to the school even if I had any reasons to stay at home"—on a scale from 1 (Strongly disagree) to 4 (Strongly agree), with 1 being low resilience and 4 being high resilience
Attend tuition classes (related to schoolwork) (no = 0, yes = 1)	A binary variable that measures whether students attend tuition classes outside of regular school hours to seek additional help with their schoolwork. Coded as 0 for "no" and 1 for "yes."
Participate in summer/winter camps (no = 0, yes = 1)	A binary variable that measures whether students participate in summer or winter camps related to their academic studies. Coded as 0 for "no" and 1 for "yes."
Student self-expectation	The degree of academic expectations that a student sets for themselves is measured on a scale from 1 to 10: 1 (Dropping out of school) indicates a low expectation, and 10 (Obtain a doctoral degree) indicates a high expectation
Student self-confidence	The measure of a student's self-perceived academic confidence is rated on a 4-point scale, ranging from 1 (Not confident at all) to 4 (Very confident), with higher scores indicating greater confidence and lower scores indicating lower confidence
Attend International Mathematical Olympiad (IMO) class	A binary variable that measures whether the student attends IMO class: 1 indicates attendance, and 0 indicates nonattendance
Attend extra Mathematics (exclude IMO) class	A binary variable that measures whether the student attends extra Mathematics (exclude IMO) class, with a value of 1 indicating attendance and 0 indicating nonattendance
Attend extra Chinese class	A binary variable that measures whether the student attends extra Chinese class: 1 indicates attendance, and 0 indicates nonattendance
Attend extra English class	A binary variable that measures whether the student attends extra English class: 1 indicates attendance, and 0 indicates nonattendance
Frequency of visits to museums	How frequently the student visits museums, ranging from 1, the least frequent, to 6, the most frequent
Parental effort	
Help with their child's homework	The amount of help parents provide to their child with their homework on a scale from 1 (There is no need to help) to 5 (Yes, help is provided almost every day)
Supervise their child's homework	The level of supervision parents provide to their child while they complete their homework on a scale of 0 (Not at all) to 4 (Completely)
Frequency of parental visits to museums with their Child's	The number of visits to museums by parents with their child on a scale from 1 (Never) to 6 (More Than Once a Week)
Parents talk to teachers about their child's learning	A binary variable that measures whether parents communicate with their child's teachers about their learning: 0 (No) to 1 (Yes)
Whether parents proactively contact teachers	The level of proactivity demonstrated by parents in contacting their child's teachers on a scale from 1 (Never) to 4 (five times or more)

Table 3 (continued)

Variables	Operationalizations
Parental discipline for their child	A binary variable that measures the level of discipline imposed by parents on their child on a scale of 0 (no discipline) to 1 (high discipline)
Parents enrolling their Childs in tuition	A binary variable that measures on a scale from 0 (no enrollment) to 1 (enrollment) whether parents enroll their child in tuition classes
Parents' concern for their child's effort level	A binary variable that measures on a scale from 0 (low concern) to 1 (high concern) the level of concern parents have for their child's effort in their academic work
Parents' requirements for their child's performance	The level, on a scale from 1 (No special requirement) to 4 (Being one of the top five of his/her class), of the academic record parents require from their child
Parent's educational expectations for their child	The level, on a scale from 1 (Drop out now) to 9 (Obtain a doctoral degree), of educational expectations parents have for their child
Parent's faith in their child	The level, on a scale from 1 (Not confident at all) to 4 (Very confident)
Parents being strict about their child's homework and exams	The level, on a scale from 1 (I don't care) to 3 (I'm very strict about it), of strictness parents impose on their child's homework and exam performance
Parents being strict about their child's school behaviors	The level, on a scale from 1 (I don't care) to 3 (I'm very strict about it), of strictness parents impose on their child's behavior in school
Parents being strict about their child's time spent on the internet	The level, on a scale from 1 (I don't care) to 3 (I'm very strict about it), of strictness parents impose on their child's time spent on the internet
Parents being strict about their child's time spent on TV	The level, on a scale from 1 (I don't care) to 3 (I'm very strict about it), of strictness parents impose on their child's time spent watching TV
School effort	
School requires students to attend night study	A binary variable that measures whether the school requires students to attend night study or not; the scale is 1 (No), 2 (Yes, Grade nine only), and 3 (Yes, Grade seven and Grade nine)
Teachers on duty for night study	A binary variable that measures the teachers' involvement in night study: 0 (No) and 1 (Yes)
School organizes summer/winter camps for students	A binary variable that measures whether the school organizes summer/winter camps for students: 0 (No) and 1 (Yes)
Frequency of school sessions on academic/life coaching	The frequency of school sessions on academic/life coaching, with 1 (Never) indicating low frequency and 4 (Over five times) indicating high frequency
Availability of teacher training	A binary variable that measures whether the school provides training programs for teachers: 0 (No) and 1 (Yes)
Partnerships with local businesses for additional resources	A binary variable that measures whether a school has partnerships with local businesses and organizations for additional resources for students: 0 (No) and 1 (Yes)
Frequency of parent-teacher meetings	The frequency of parent-teacher meetings, with 1 (Never) indicating low frequency and 4 (Over five times) indicating high frequency
Frequency of written reports from the school to parents	The frequency of written reports from the school to parents, with 1 (Never) indicating low frequency and 4 (Over five times) indicating high frequency
Frequency of schools inviting parents to observe	The frequency of schools inviting parents to observe, with 1 (Never) indicating low frequency and 4 (Over five times) indicating high frequency
Main teaching methods: teacher-led lectures	A binary variable that measures whether the school's main teaching method is teacher-led lectures: 0 (No) and 1 (Yes)
Main teaching methods: group discussions	A binary variable that measures whether the school's main teaching method is group discussions: 0 (No) and 1 (Yes)
Main teaching methods: bilingual teaching	A binary variable that measures whether the school's main teaching method is bilingual teaching: 0 (No) and 1 (Yes)
Main teaching methods: stratified teaching	A binary variable that measures whether the school's main teaching method is stratified teaching: 0 (No) and 1 (Yes)
School takes remedial classes for students with failing grades	The level of remedial classes offered to students with failing grades, with 0 indicating that no remedial classes are offered and 4 indicating a high level of remedial classes

Table 3 (continued)

Variables	Operationalizations
School offers advanced study for students good at a single subject	A binary variable that measures whether the school offers further enhanced opportunities to outstanding students: 0 (No) and 1 (Yes)
Controls	
Gender (male = 0, female = 1)	A binary variable that indicates students' gender, coded as 0 for male and 1 for female
House registration type (rural = 0, urban = 1)	A binary variable that indicates students' house registration type, coded as 0 for rural and 1 for urban
Cognitive ability	An ordinal variable that indicates students' cognitive ability, ranging from 0 to 35
Family structure (only-child = 0, non-only child = 1)	A binary variable that indicates students' family structure, coded as 0 for only-child and 1 for non-only child
Health condition	An ordinal variable that indicates students' health condition, using a rating scale of 1 to 5, with 1 being very poor health and 5 being very good health
Parent's educational level	This variable is measured by asking the parents about their educational qualifications and coding them as follows: 1 for None, 2 for Finished elementary school, 3 for Junior high school degree, 4 for Technical secondary school or technical school, 5 for Vocational high school degree, 6 for Senior high school degree, 7 for Junior college degree, 8 for Bachelor degree, and 9 for Master degree or higher
Parent's occupation	This variable is measured by asking the parents about their occupation and coding them as follows: 1 for unskilled worker, 2 for skilled worker, 3 for clerical or sales, 4 for service worker, 5 for small business owner, 6 for professional, 7 for executive or managerial, 8 for retired, 9 for unemployed, 10 for student, 11 for homemaker, 12 for farmer, 13 for other
Parent's income	This variable is measured by asking the parents about their income level and coding it on a scale from 1 (Very poor) to 5 (Very rich)
Separate studying desk available (no = 0, yes = 1)	A binary variable that measures whether a family has a separate studying desk for the child: 0 (No) and 1 (Yes)
Computer and internet available (no = 0, yes = 1)	A binary variable that measures whether a family has computer and internet access: 0 (No) and 1 (Yes)
School ranking	This variable is measured using a scale from 1 to 5, with 1 (Near the bottom) being the lowest-ranked school and 5 (Among the best) being the highest-ranked school
School category (private = 0, public = 1)	A binary variable that measures whether the school is 0 (private-funded) or 1 (public-funded)
School size	This variable is measured by the number of classrooms owned in the school, ranging from 5 to 400
Student-teacher ratio (teacher = 1)	This variable is measured by the number of students per teacher (teacher = 1), ranging from 3 to 30.80
School fiscal per year	This variable is measured by the amount of money allocated for the school's operation per year, ranging from 0 to 100

Control variables

Student demographic characteristics To control for the potential influence of student demographic characteristics, I include 5 student demographic variables as control variables: gender (male = 0, female = 1), family structure (only child = 0, non-only child = 1), health condition (rated on a scale from 1 = very poor to 5 = very good), and house registration type (rural = 0, urban = 1). Table 4 displays the distribution of these variables in the sample, with male

students comprising 52% of the sample and female students comprising 48%. Of the students, 56% reported having siblings, and 41% reported being healthy. In addition to demographic variables, cognitive ability is an important factor in academic success. The CEPS used a 35-question cognitive ability test administered to junior high school students to measure cognitive ability. On average, the students correctly answered 14 of the 35 questions. Higher scores on the test indicate higher cognitive ability levels.

Table 4 Descriptive statistics for controlled variables

Controls						
Individual demographic characteristics						
Gender (male = 0, female = 1)	StuGender	0.486	0.500	0	1	
House registration type (rural = 0, urban = 1)	StuHukou	0.458	0.498	0	1	
Cognitive ability	StuCognition	13.90	8.134	0	35	
Family structure (only-child = 0, non-only child = 1)	StuOnlyChild	0.563	0.496	0	1	
Health condition	StuHealth	4.144	0.866	1	5	
Family background characteristics						
Parent's educational level	ParEducation	4.075	1.995	1	9	
Parent's occupation	ParOccupation	6.543	2.614	1	14	
Parent's income	ParIncome	2.809	0.601	1	5	
Separate studying desk available (no = 0, yes = 1)	StuOwnDesk	0.789	0.408	0	1	
Computer and internet available (no = 0, yes = 1)	StuOwnInternet	0.620	0.485	0	1	
School resource characteristics						
School ranking	SchRanking	3.948	0.845	1	5	
School category (private = 0, public = 1)	SchCategory	0.923	0.266	0	1	
School size	SchSize	40.62	40.55	5	400	
Student-teacher ratio (teacher = 1)	SchTeacherRatio	13.33	4.537	3	30.8	
School fiscal per year	SchFiscalYear	33.17	42.63	0	100	
<i>N</i>		24,974				

Family background characteristics Family background, often reflected in parents' income, professions, and educational levels, is a crucial predictor of academic achievement and potentially affects students' and parental efforts. To control for these factors, I collect data from 5 answers from parents' questionnaires. As shown in Table 4, 72% of moderately affluent families, 40% of parents have completed at least a junior high school degree, and 53% of parents have obtained higher-level occupations such as technical worker, teacher, engineer, doctor, lawyer, and government official. In addition, I also consider the availability of a separate studying desk, computer, and internet at home because limited access to these resources can create inequities in academic opportunities and outcomes. Descriptive statistics indicate that approximately 79% of students have a separate studying desk, and 62% are equipped with a computer and internet at home, potentially facilitating their learning process.

School resource characteristics To account for the influence of school characteristics on academic achievement and effort levels, I include 5 variables related to the school's resources. One such variable is school ranking, which is determined based on academic performance and is represented by higher values indicating better performance. Of the total sample, approximately 79% of schools have above-average rankings and are considered among the best. Regarding school categories, 92% of schools are public-funded, and the remaining 7.47% are private schools. Additionally, the size of schools is considered, with approximately 80% having a below-middle size that

accommodates 50 or fewer classrooms. The teacher-to-students ratio is also a factor; 26% of schools have adequate teachers, with a teacher-to-students ratio of less than one: ten (one teacher for ten students). Last, I include fiscal resources available to schools per year, with 55% of schools eligible to receive provincial-level funding annually.

Acknowledgements The author would like to express her heartfelt gratitude to Prof. Eva Jablonka and Dr. Yulin Hao for their unwavering support, valuable guidance, and constant encouragement throughout the development of this manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. The author acknowledges support by the Open Access Publication Fund of Freie Universität Berlin.

Declarations

Competing interests The author confirms that there are no conflict of interest or significant financial support that could influence the outcome of the submitted work.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., & Venkatasubramanian, S. (2018). Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54, 95–122. <https://doi.org/10.1007/s10115-017-1116-3>
- Al-Samarraie, H., & Saeed, N. (2018). A systematic review of cloud computing tools for collaborative learning: Opportunities and challenges to the blended-learning environment. *Computers & Education*, 124, 77–91. <https://doi.org/10.1016/j.compedu.2018.05.016>
- Amis, J. M., Mair, J., & Munir, K. A. (2020). The organizational reproduction of inequality. *Academy of Management Annals*, 14(1), 195–230. <https://doi.org/10.5465/annals.2017.0033>
- Avvisati, F., Besbas, B., & Guyon, N. (2010). Parental involvement in school: A literature review. *Revue D'économie Politique*, 5, 759–778.
- Baños, J. H., Noah, J. P., & Harada, C. N. (2019). Predictors of student engagement in learning communities. *Journal of Medical Education and Curricular Development*, 6, 2382120519840330. <https://doi.org/10.1177/2382120519840330>
- Baumrind, D. (1971). Current patterns of parental authority. *Developmental Psychology*, 4(12), 1.
- Berkowitz, R., Moore, H., Astor, R. A., & Benbenishty, R. (2017). A research synthesis of the associations between socioeconomic background, inequality, school climate, and academic achievement. *Review of Educational Research*, 87(2), 425–469. <https://doi.org/10.3102/0034654316669821>
- Blau, P. M., & Duncan, O. D. (1967). The American occupational structure. <https://www.journals.uchicago.edu/doi/pdf/https://doi.org/10.1086/224796>
- Boonk, L., Gijsselaers, H. J., Ritzen, H., & Brand-Gruwel, S. (2018). A review of the relationship between parental involvement indicators and academic achievement. *Educational Research Review*, 24, 10–30. <https://doi.org/10.1016/j.edurev.2018.02.001>
- Bourdieu, P. (2002). The social conditions of the international circulation of ideas. *Actes De La Recherche En Sciences Sociales*, 5, 3–8. <https://doi.org/10.3917/arss.145.0003>
- Bourdieu, P. (2014). The habitus and the space of life-styles (1984). *The people, place, and space reader* (pp. 173–178). Routledge.
- Broer, M., Bai, Y., & Fonseca, F. (2019). *Socioeconomic inequality and educational outcomes: Evidence from twenty years of TIMSS* (p. 83). Springer.
- Burić, I., & Sorić, I. (2012). The role of test hope and hopelessness in self-regulated learning: Relations between volitional strategies, cognitive appraisals and academic achievement. *Learning and Individual Differences*, 22(4), 523–529. <https://doi.org/10.1016/j.lindif.2012.03.011>
- Caldas, S. J., & Bankston, C. (1997). Effect of school population socioeconomic status on individual academic achievement. *The Journal of Educational Research*, 90(5), 269–277. <https://doi.org/10.1080/00220671.1997.10544583>
- Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology*, 62, 583–619. <https://doi.org/10.1146/annurev.psych.093008.100356>
- Dalalyan, A. S., Hebiri, M., & Lederer, J. (2017). On the prediction performance of the lasso. *Bernoulli*, 23(1), 552–581. <https://doi.org/10.3150/15-BEJ756>
- Darling-Hammond, L., & Cook-Harvey, C. M. (2018). Educating the whole child: Improving school climate to support student success. *Learning Policy Institute*. <https://learningpolicyinstitute.org/product/educating-whole-child>.
- Darling-Hammond, L. (2018). From “separate but equal” to “No Child Left Behind”: The collision of new standards and old inequalities. In E. B. Hilty & E. B. Hilty (Eds.), *Thinking about schools* (pp. 419–437). Routledge.
- De Fraja, G., Oliveira, T., & Zanchi, L. (2010). Must try harder: Evaluating the role of effort in educational attainment. *The Review of Economics and Statistics*, 92(3), 577–597.
- Deluca, S., & Rosenbaum, J. E. (2001). Individual agency and the life course: Do low-SES students get less long-term payoff for their school efforts? *Sociological Focus*, 34(4), 357–376. <https://doi.org/10.1080/00380237.2001.10571208>
- Deming, D. J., Hastings, J. S., Kane, T. J., & Staiger, D. O. (2014). School choice, school quality, and postsecondary attainment. *American Economic Review*, 104(3), 991–1013. <https://doi.org/10.1257/aer.104.3.991>
- Dietrich, H., Patzina, A., & Lerche, A. (2021). Social inequality in the homeschooling efforts of German high school students during a school closing period. *European Societies*, 23(sup1), S348–S369. <https://doi.org/10.1080/14616696.2020.1826556?needAccess=true>
- Dunlosky, J., Badali, S., Rivers, M. L., & Rawson, K. A. (2020). The role of effort in understanding educational achievement: Objective effort as an explanatory construct versus effort as a student perception. *Educational Psychology Review*, 32, 1163–1175. <https://doi.org/10.1007/s10648-020-09577-3>
- Đuričić, M., & Bunijevac, M. (2017). Parental involvement as a important factor for successful education. *Center for Educational Policy Studies Journal*, 7(3), 137–153. <https://doi.org/10.26529/cepsj.291>
- Edmark, K., & Persson, L. (2021). The impact of attending an independent upper secondary school: Evidence from Sweden using school ranking data. *Economics of Education Review*, 84, 102148. <https://doi.org/10.1016/j.econedurev.2021.102148>
- Edossa, A. K., Schroeders, U., Weinert, S., & Artelt, C. (2018). The development of emotional and behavioral self-regulation and their effects on academic achievement in childhood. *International Journal of Behavioral Development*, 42(2), 192–202. <https://doi.org/10.1177/0165025416687412>
- Eryong, X., & Li, J. (2022). What is the value essence of “double reduction” (Shuang Jian) policy in China? A policy narrative perspective. *Educational Philosophy and Theory*. <https://doi.org/10.1080/00131857.2022.2040481>
- Fang, S., Huang, J., Curley, J., & Birkenmaier, J. (2018). Family assets, parental expectations, and children educational performance: An empirical examination from China. *Children and Youth Services Review*, 87, 60–68. <https://doi.org/10.1016/j.childyouth.2018.02.018>
- Fu, G. (2020). The knowledge-based versus student-centred debate on quality education: Controversy in China’s curriculum reform. *Compare: A Journal of Comparative and International Education*, 50(3), 410–427. <https://doi.org/10.1080/03057925.2018.1523002>
- Gamboa, L. F., & Waltenberg, F. D. (2012). Inequality of opportunity for educational achievement in Latin America: Evidence from PISA 2006–2009. *Economics of Education Review*, 31(5), 694–708. <https://doi.org/10.1016/j.econedurev.2012.05.002>
- Ganzeboom, H. B., Treiman, D. J., & Ultee, W. C. (1991). Comparative intergenerational stratification research: Three generations and beyond. *Annual Review of Sociology*, 17(1), 277–302. <https://doi.org/10.1146/annurev.so.17.080191.001425>
- Gbollie, C., & Keamu, H. P. (2017). Student academic performance: The role of motivation, strategies, and perceived factors hindering Liberian junior and senior high school students learning. *Education Research International*. <https://doi.org/10.1155/2017/1789084>
- Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights*, 1(3), 291–308. <https://doi.org/10.1257/aeri.20180633>

- Golley, J., & Kong, S. T. (2018). Inequality of opportunity in China's educational outcomes. *China Economic Review*, *51*, 116–128. <https://doi.org/10.1016/j.chieco.2016.07.002>
- Guan, J., Xiang, P., McBride, R., & Bruene, A. (2006). Achievement goals, social goals, and students' reported persistence and effort in high school physical education. *Journal of Teaching in Physical Education*, *25*(1), 58–74. <https://doi.org/10.1123/jtpe.25.1.58>
- Hirsch, E. D. (2019). *Why knowledge matters: Rescuing our children from failed educational theories*. Harvard Education Press.
- Hodge, B., Wright, B., & Bennett, P. (2018). The role of grit in determining engagement and academic outcomes for university students. *Research in Higher Education*, *59*, 448–460. <https://doi.org/10.1007/s11162-017-9474-y>
- Hong, J. C., Hwang, M. Y., Tai, K. H., & Lin, P. H. (2017). Intrinsic motivation of Chinese learning in predicting online learning self-efficacy and flow experience relevant to students' learning progress. *Computer Assisted Language Learning*, *30*(6), 552–574. <https://doi.org/10.1080/09588221.2017.1329215>
- Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The effort paradox: Effort is both costly and valued. *Trends in Cognitive Sciences*, *22*(4), 337–349. <https://doi.org/10.1016/j.tics.2018.01.007>
- Kohn, M. L., Naoi, A., Schoenbach, C., Schooler, C., & Slomczynski, K. M. (1990). Position in the class structure and psychological functioning in the United States, Japan, and Poland. *American Journal of Sociology*, *95*(4), 964–1008. <https://doi.org/10.1086/229382>
- Lareau, A. (2002). Invisible inequality: Social class and childrearing in black families and white families. *American Sociological Review*. <https://doi.org/10.2307/3088916>
- Ling, M. (2017). Precious son, reliable daughter: Redefining son preference and parent–child relations in migrant households in urban China. *The China Quarterly*, *229*, 150–171. <https://doi.org/10.1017/S0305741016001570>
- Ma, G., & Wu, Q. (2019). Social capital and educational inequality of migrant children in contemporary China: A multilevel mediation analysis. *Children and Youth Services Review*, *99*, 165–171. <https://doi.org/10.1016/j.childyouth.2019.02.002>
- Masci, C., Johnes, G., & Agasisti, T. (2018). Student and school performance across countries: A machine learning approach. *European Journal of Operational Research*, *269*(3), 1072–1085.
- Mudrak, J., Slepicka, P., Slepickova, I., Zabrodská, K., & Knoblochová, M. (2021). Motivational beliefs and subjective effort in adolescent athletes. *International Journal of Sport Psychology*, *52*, 335–354. <https://doi.org/10.7352/IJSP.2021.52.335>
- Ng, F. F. Y., & Wei, J. (2020). Delving into the minds of Chinese parents: What beliefs motivate their learning-related practices? *Child Development Perspectives*, *14*(1), 61–67. <https://doi.org/10.1111/cdep.12358>
- Ogutu, J. O., Schulz-Streeck, T., & Piepho, H. P. (2012). Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions. In BMC proceedings. *BioMed Central*, *6*(2), 1–6.
- Park, S., Stone, S. I., & Holloway, S. D. (2017). School-based parental involvement as a predictor of achievement and school learning environment: An elementary school-level analysis. *Children and Youth Services Review*, *82*, 195–206. <https://doi.org/10.1016/j.childyouth.2017.09.012>
- Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353. <https://doi.org/10.1037/a0026838>
- Roick, J., & Ringeisen, T. (2017). Self-efficacy, test anxiety, and academic success: A longitudinal validation. *International Journal of Educational Research*, *83*, 84–93. <https://doi.org/10.1016/j.ijer.2016.12.006>
- Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, *61*, 101860. <https://doi.org/10.1016/j.cedpsych.2020.101860>
- Schunk, D. H., & DiBenedetto, M. K. (2020). Motivation and social cognitive theory. *Contemporary Educational Psychology*, *60*, 101832. <https://doi.org/10.1016/j.cedpsych.2019.101832>
- Sewell, W. H., & Shah, V. P. (1968). Social class, parental encouragement, and educational aspirations. *American Journal of Sociology*, *73*(5), 559–572. <https://doi.org/10.1086/224530>
- Sheldon, S. B., & Epstein, J. L. (2005). Involvement counts: Family and community partnerships and mathematics achievement. *The Journal of Educational Research*, *98*(4), 196–207. <https://doi.org/10.3200/JOER.98.4.196-207>
- Shenhav, A., Prater Fahey, M., & Grahek, I. (2021). Decomposing the motivation to exert mental effort. *Current Directions in Psychological Science*, *30*(4), 307–314. <https://doi.org/10.1177/09637214211009510>
- Soni, A., & Kumari, S. (2017). The role of parental math anxiety and math attitude in their children's math achievement. *International Journal of Science and Mathematics Education*, *15*, 331–347. <https://doi.org/10.1007/s10763-015-9687-5>
- Stables, A., Murakami, K., McIntosh, S., & Martin, S. (2014). Conceptions of effort among students, teachers and parents within an English secondary school. *Research Papers in Education*, *29*(5), 626–648. <https://doi.org/10.1080/02671522.2013.878376>
- Steele, J. (2020). What is (perception of) effort? Objective and subjective effort during task performance. *PsyArXiv*. <https://doi.org/10.31234/osf.io/kbyhm>
- Steyerberg, E. W., & Harrell, F. E., Jr. (2016). Prediction models need appropriate internal, internal-external, and external validation. *Journal of Clinical Epidemiology*, *69*, 245. <https://doi.org/10.1016/j.jclinepi.2015.04.005>
- Takeda, A., Niranjana, M., Gotoh, J. Y., & Kawahara, Y. (2013). Simultaneous pursuit of out-of-sample performance and sparsity in index tracking portfolios. *Computational Management Science*, *10*(1), 21–49. <https://doi.org/10.1007/s10287-012-0158-y>
- Tang, L., & Horta, H. (2021). Women academics in Chinese universities: A historical perspective. *Higher Education*. <https://doi.org/10.1007/s10734-020-00669-1>
- Verge, T. (2021). Gender equality policy and universities: Feminist strategic alliances to re-gender the curriculum. *Journal of Women, Politics & Policy*, *42*(3), 191–206. <https://doi.org/10.1080/1554477X.2021.1904763>
- Wang, W. (2016). China education panel survey baseline report. From <https://opendata.pku.edu.cn/dataset.xhtml?persistentId=doi:https://doi.org/10.18170/DVN/KURJUJ>
- Wang, J., Tigelaar, D. E., & Admiraal, W. (2019). Connecting rural schools to quality education: Rural teachers' use of digital educational resources. *Computers in Human Behavior*, *101*, 68–76. <https://doi.org/10.1016/j.chb.2019.07.009>
- Winship, C. (1992). Race, poverty, and The American Occupational Structure [Review of The American Occupational Structure, by P. Blau & O. D. Duncan]. *Contemporary Sociology*, *21*(5), 639–643. <https://doi.org/10.2307/2075545>
- Won, S., Wolters, C. A., & Mueller, S. A. (2018). Sense of belonging and self-regulated learning: Testing achievement goals as

- mediators. *The Journal of Experimental Education*, 86(3), 402–418. <https://doi.org/10.5204/ssj.v8i2.376z>
- Xu, D. (2016). Three essays on educational inequality in China: the causal impacts of migration, segregation, and college education (Doctoral dissertation). <http://hdl.handle.net/1783.1/95298>
- Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist*, 47(4), 302–314. <https://doi.org/10.1080/00461520.2012.722805>
- Zhu, Y. (2019). New national initiatives of modernizing education in China. *ECNU Review of Education*, 2(3), 353–362. <https://doi.org/10.1177/2096531119868069>
- Zimmerman, B. J. (2013). From cognitive modeling to self-regulation: A social cognitive career path. *Educational Psychologist*, 48(3), 135–147. <https://doi.org/10.1080/00461520.2013.794676>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.