

Journal of Educational Psychology

How the Predictors of Math Achievement Change Over Time: A Longitudinal Machine Learning Approach

Rosa Lavelle-Hill, Anne C. Frenzel, Thomas Goetz, Stephanie Lichtenfeld, Herbert W. Marsh, Reinhard Pekrun, Michiko Sakaki, Gavin Smith, and Kou Murayama

Online First Publication, September 5, 2024. <https://dx.doi.org/10.1037/edu0000863>

CITATION

Lavelle-Hill, R., Frenzel, A. C., Goetz, T., Lichtenfeld, S., Marsh, H. W., Pekrun, R., Sakaki, M., Smith, G., & Murayama, K. (2024). How the predictors of math achievement change over time: A longitudinal machine learning approach. *Journal of Educational Psychology*. Advance online publication. <https://dx.doi.org/10.1037/edu0000863>

How the Predictors of Math Achievement Change Over Time: A Longitudinal Machine Learning Approach

Rosa Lavelle-Hill^{1, 2, 3}, Anne C. Frenzel⁴, Thomas Goetz⁵, Stephanie Lichtenfeld⁶, Herbert W. Marsh⁷,
Reinhard Pekrun^{4, 7, 8}, Michiko Sakaki^{1, 9}, Gavin Smith¹⁰, and Kou Murayama^{1, 9}

¹ Hector Research Institute of Education Sciences and Psychology, University of Tübingen

² Department of Psychology, University of Copenhagen

³ Copenhagen Center for Social Data Science (SODAS), University of Copenhagen

⁴ Department of Psychology, University of Essex

⁵ Department of Developmental and Educational Psychology, Faculty of Psychology, University of Vienna

⁶ Educational Psychology, Faculty of Education, University of Hamburg

⁷ Institute for Positive Psychology and Education, Australian Catholic University

⁸ Department of Psychology, Ludwig-Maximilians-Universität München

⁹ Research Institute, Kochi University of Technology

¹⁰ N/LAB, Business School, University of Nottingham



Researchers have focused extensively on understanding the factors influencing students' academic achievement over time. However, existing longitudinal studies have often examined only a limited number of predictors at one time, leaving gaps in our knowledge about how these predictors collectively contribute to achievement beyond prior performance and how their impact evolves during students' development. To address this, we employed machine learning to analyze longitudinal survey data from 3,425 German secondary school students spanning 5 to 9 years. Our objectives were twofold: to model and compare the predictive capabilities of 105 predictors on math achievement and to track changes in their importance over time. We first predicted standardized math achievement scores in Years 6–9 using the variables assessed in the previous year (“next year prediction”). Second, we examined the utility of the variables assessed in Year 5 at predicting future math achievement at varying time lags (1–4 years ahead)—“varying lag prediction.” In the next year prediction analysis, prior math achievement was the strongest predictor, gaining importance over time. In the varying lag prediction analysis, the predictive power of Year 5 math achievement waned with longer time lags. In both analyses, additional predictors, including intelligence quotient, grades, motivation and emotion, cognitive strategies, classroom/home environments, and demographics (including socioeconomic status), exhibited relatively smaller yet consistent contributions, underscoring their distinct roles in predicting math achievement over time. The findings have implications for both future research and educational practices, which are discussed in detail.

Samuel Greiff served as action editor.

Rosa Lavelle-Hill  <https://orcid.org/0000-0002-1767-9828>


This research was supported by the Alexander von Humboldt Foundation (the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research) to Kou Murayama. The PALMA study was supported by grants from the German Research Foundation (Deutsche Forschungsgemeinschaft) awarded to Reinhard Pekrun.

Anne C. Frenzel, Thomas Goetz, Stephanie Lichtenfeld, Herbert W. Marsh, Reinhard Pekrun, Michiko Sakaki, and Gavin Smith are arranged in alphabetical order. The authors have no conflicts of interest to disclose.

This work is licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International License (CC BY-NC-ND 4.0; <https://creativecommons.org/licenses/by-nc-nd/4.0>). This license permits copying and redistributing the work in any medium or format for noncommercial use provided the original authors and source are credited and a link to the license is included in attribution. No derivative works are permitted under this license.

Rosa Lavelle-Hill served as lead for formal analysis, methodology, visualization, writing—original draft, and writing—review and editing and contributed equally to conceptualization. Anne C. Frenzel served in a supporting role for writing—original draft and writing—review and editing. Thomas Goetz served

in a supporting role for writing—review and editing. Stephanie Lichtenfeld served in a supporting role for writing—original draft and writing—review and editing. Herbert W. Marsh served in a supporting role for writing—review and editing. Reinhard Pekrun contributed equally to funding acquisition and served in a supporting role for project administration, writing—original draft, and writing—review and editing. Michiko Sakaki served in a supporting role for writing—original draft and writing—review and editing. Gavin Smith served in a supporting role for methodology. Kou Murayama served as lead for conceptualization, funding acquisition, and supervision, contributed equally to methodology, writing—original draft, and writing—review and editing, and served in a supporting role for data curation. Anne C. Frenzel, Thomas Goetz, Stephanie Lichtenfeld, and Reinhard Pekrun contributed equally to data curation. Anne C. Frenzel, Thomas Goetz, and Stephanie Lichtenfeld contributed equally to project administration.

 The study materials are available at <https://github.com/Rosa-Lavelle-Hill/palma-ml-open>.

Correspondence concerning this article should be addressed to Rosa Lavelle-Hill, Copenhagen Center for Social Data Science (SODAS), University of Copenhagen, Øster Farimagsgade 2A, 1353 København, Denmark; Department of Psychology, University of Copenhagen, Øster Farimagsgade 2A, 1353 København, Denmark. Email: rla@psy.ku.dk

Educational Impact and Implications Statement

Understanding the predictors of students' academic achievement is one of the foremost concerns in research on education. Most studies analyze the effects of only a handful of predictors at one time. However, in the real world, many factors likely interact and jointly contribute to explaining achievement. We use machine learning methods to model a large number of variables and their interactions to better understand how accurately data collected from school documents, cognitive tests, and self-report questionnaires can predict students' math achievement, above and beyond prior achievement. We also assess how the predictive utility of groups of variables changes over time. The insights produced are useful for understanding what data are most useful to collect when predicting math achievement, as well as when to plan interventions to be maximally effective.

Keywords: mathematics, student achievement, longitudinal survey data, machine learning, explainable artificial intelligence

Supplemental materials: <https://doi.org/10.1037/edu0000863.supp>

Understanding the predictors of students' academic achievement has been one of the foremost concerns in research on education. Using longitudinal panel data, research has identified numerous factors that predict math achievement over time: basic cognitive abilities (Deary et al., 2007), learning/cognitive strategies (e.g., Muis et al., 2018), motivation (e.g., Steinmayr & Spinath, 2009), the influence of teachers and schools (Hattie, 2008) as well as demographic characteristics (e.g., Howard, 2019). However, most previous studies focused on a selection of variables to predict math achievement, and only a limited number of studies have examined how these factors jointly predict achievement outcomes. More importantly, we do not have a clear idea about how these factors jointly predict future math achievement above and beyond current achievement scores (i.e., predicting change in math achievement). In addition, little is known about whether the predictive power of these factors changes over time. In this study, we employ a machine learning approach to analyze the entire set of variables from a large longitudinal panel study to address these questions, especially focusing on the development-dependent change in these factors' importance in predicting future math achievement scores.

Factors Associated With Longitudinal Change in Math Achievement

Existing studies looking at the predictors of achievement have most commonly used longitudinal panel designs; that is, they assessed many students or teachers at several time points with relatively long time intervals in between (Blossfeld et al., 2009). Longitudinal designs have considerable advantages over cross-sectional designs. In research on student achievement, they allow researchers to examine factors predicting longitudinal change in achievement scores. More specifically, the resulting data allow researchers to analyze whether and how variables can predict achievement scores in the future after controlling for baseline achievement. Such predictive analyses are often performed using sophisticated statistical models such as latent cross-lagged panel models or variations of them (for a review, see Usami et al., 2019).

Using these statistical methods, previous studies have identified many factors that predict change in math achievement over time. Studies repeatedly found motivational and emotional factors to be important predictors (e.g., Pekrun et al., 2017). For example,

Steinmayr and Spinath (2009) showed that some motivation variables (e.g., mastery goals and perceived control) predicted students' math achievement after controlling for prior achievement. In addition, research has found that cognitive strategies (or learning strategies) make a difference in students' subsequent math achievement scores (e.g., Muis et al., 2018; Murayama et al., 2013). For example, Murayama et al. (2013) showed deep processing learning strategies positively predicted change in math achievement over 3 years. Researchers have also demonstrated that family context (such as parental involvement) and class context (such as teachers' behavior) play critical roles in the development of students' math achievement (Hong & Ho, 2005; Kunter et al., 2013; Murayama et al., 2016). For example, using cross-lagged panel modeling, Hong et al. (2010) showed that reported parental value in math had positive lagged effects on math achievement in adolescents.

These findings are encouraging, as they indicate that teachers and parents can support students to perform better in the future. These studies, however, usually focus on the statistical significance of the effects from a small set of predictors, resulting in relatively small overall effect sizes. For example, Talsma et al. (2018) conducted a meta-analysis of cross-lagged effects of self-efficacy on academic achievement (predominantly math achievement). They found that the averaged standardized cross-lagged effects were .085 (short lag) and .057 (long lag). Compared to other cross-lagged effects, these are deemed "medium-sized" (Orth et al., 2024), but practically speaking, they are still relatively small. This is understandable and to be expected—it is unlikely that change in math achievement is reducible to a single factor. However, an important question is the extent to which various factors jointly (and including their interactions) influence future math achievement above and beyond current achievement. In other words, do these factors have substantive collective effects on change in math achievement?

In addition to looking at only a handful of possible predictors at one time, previous work in this area has also tended to be limited in the following two ways. Firstly, the focus is typically on grades (rather than standardized test scores) as an outcome measure of achievement, which can vary in how they are assessed across schools. Secondly, previous studies have not paid explicit attention to how the effects of predictor variables change over time, leaving it unclear when these factors are most important in the child's development. However, for those considering educational interventions,

this is important. For example, is there a critical period when educational interventions should be conducted to improve students' math achievement effectively?

There are, however, a few studies that can provide a clue to these questions. For example, Bailey et al. (2014) demonstrated that for longitudinal math achievement scores, a large proportion of the variance (i.e., 55%) is explained by latent trait effects (stable factors that influence an individual's mathematics achievement similarly over the course of development), and unstable components (states), such as the influences of specific teachers, explain less variance (i.e., 7%). These results indicate that many of the less stable factors identified above (i.e., cognitive strategies, emotions, and classroom context) may have a limited role in changes to achievement. Other studies have examined how achievement changes during development, and generally suggest that the relationship between prior achievement and current achievement becomes stronger over the school years (Geary et al., 2017; K. Lee & Bull, 2016; Lin & Powell, 2022). For example, Geary et al. (2017) showed with longitudinal data from 167 children that the effect of prior math achievement on subsequent math achievement (1 year apart) constantly increased from Grade 2 to Grade 8. These results suggest the possibility that the longer students study at school, the less space for other factors to explain additional changes in math achievement. However, these studies analyzing the stability of achievement over time did not include other critical factors that have been shown to influence math achievement scores (e.g., motivational and emotional variables, classroom context, etc.) and did not assess how all these factors collectively predict math achievement scores beyond prior achievement, nor how the predictive effects change over the child's development.

Leveraging Machine Learning Methods to Examine the Predictors of Longitudinal Change in Math Achievement

To address some of the limitations of previous studies, the current paper takes a different approach: Analyzing all of the variables included in a large longitudinal data set to predict math achievement over time, which we coin an "all-inclusive approach" (see also Tamura et al., 2022). In the real world, numerous factors exist that are likely to predict achievement simultaneously. Furthermore, some predictors may interact, and their effects can even cancel each other out. Therefore, to get a realistic picture of the joint predictors of academic achievement, the entirety of the set of potential predictors, along with their nonlinear and complex interactions, should ideally be modeled at once. Such a comprehensive investigation poses a challenge to traditional statistical analysis methods due to multicollinearity issues and the high likelihood of overfitting (i.e., fitting a complex model to a sample of data that will not generalize beyond that sample). However, the application of machine learning methods in this context makes such an all-inclusive approach possible while also helps to ensure the results are generalizable to new data through processes such as cross-validation and out-of-sample testing (Hastie et al., 2009; Strobl et al., 2009).

Statistical machine learning, developed in computer science, is a methodology that allows for simultaneously detecting complex (nonlinear) patterns in data and helping to ensure that the results are not caused by idiosyncratic characteristics of the sample data (i.e., randomness, referred to as "noise" in machine learning), and thus are generalizable to new data with the same distributional

properties (Hastie et al., 2009; Yarkoni & Westfall, 2017). A typical (so-called "supervised") machine learning methodology involves the detection of relationships between a set of predictor variables (also called features) and an outcome variable in training data via repeatedly fitting and evaluating different models using a cross-validation procedure¹ and then evaluating the best-performing model on a separate "test" data set (i.e., out-of-sample testing). Importantly, this out-of-sample testing procedure, as well as certain mechanisms in machine learning models such as regularization (see the Method section), enables many different predictor variables to be entered into the model while helping to protect against overfitting. In traditional statistical methods (e.g., multiple linear regression), if there are many predictor variables, the model can become unstable due to overfitting and multicollinearity (Yarkoni & Westfall, 2017). In other words, such a model has a risk of capturing peculiar features of the fitted data, and the results may not replicate when the model is applied to new data (Babyak, 2004). Because machine learning methods always evaluate the fitted model on unseen test data, the risk of overfitting is attenuated (for more details on a typical machine learning pipeline, see Lavelle-Hill et al., 2023).

The use of machine learning methods in the social sciences has often been criticized for primarily focusing on prediction and not providing useful explanations (i.e., "black box" methods; Cox et al., 2020). To overcome this limitation, various approaches have been proposed to quantify the relative importance of the different variables in the model and their interactions (called variable importance). These variable importance measures are, however, dependent on the model and method for calculating importance and should be interpreted with caution (Henninger et al., 2023; Molnar et al., 2020; see also the General Discussion section), and ideally in combination with relevant theories (Van Lissa, 2022). Also, given that machine learning models can include many predictors that are likely to causally influence each other in different directions (i.e., have reciprocal effects), these measures are also unlikely to reflect only the causal effects of the predictors on the outcome. Nevertheless, variable importance metrics can provide important insights into the different predictor variables' relative contribution to predicting an outcome, given the model.

In educational research, machine learning methods have been most commonly used in the field of learning analytics (LA), described as "the measurement, collection, analysis, and reporting of data about learners and their contexts..." (Siemens & Long, 2011), and its sister field educational data mining (EDM), which has a greater focus on automated adaptation (e.g., task difficulty) of interactive online learning environments (Baker & Yacef, 2009). LA tends to focus on predicting academic achievement from an applied perspective. For example, LA can aim to identify students who might be at risk of dropping out or failing and to provide students with personalized or "just in time" support (Wong & Li, 2020). In addition, studies have used machine learning to predict student grades, exercise performance, and even question correctness

¹There are many different ways to perform cross-validation (e.g., leave-one-out cross-validation, temporal cross-validation, jackknife cross-validation, and nested cross-validation, see Hastie et al., 2009), but commonly K-fold cross-validation is used (see the Data Analysis section). Note that cross-validation may not be necessary with sufficient data instead, a single validation set could be used (Browne, 2000).

(Cetintas et al., 2009; Chakrapani & Chitradevi, 2022; Hellas et al., 2018; S. Lee & Chung, 2019)—some to better understand the learning process (e.g., Deininger et al., 2023).

LA research most commonly harnesses data from online learning management systems (LMSs; e.g., Moodle or Canvas) or digital learning environments such as Intelligent Tutoring Systems (but see also Goldberg et al., 2021; Gomes et al., 2013; Rawson et al., 2017). For this reason, the majority of LA research is conducted with samples from higher education where more usable data are available, for example, from LMSs (Sekeroglu et al., 2021). Such data typically consist of a combination of administrative data and fine-grained process or behavioral trace data collected from interactions with digital learning support tools or environments (Bilal et al., 2022; Daza et al., 2022; Picciano, 2012). Applying machine learning and other analytic methods to this data can help to build tools to support students, teachers, and course coordinators/designers in practice (e.g., data dashboards; Molenaar & Knoop-van Campen, 2017, Infinite Campus; Christie et al., 2019, BrightBytes; Esbenshade et al., 2023, and Civitas; Civitas, 2023).

Applied prediction models used in LA can produce useful insights for practitioners. Despite this, machine learning methods are less commonly used to inform theory in educational and developmental psychology (Dawson et al., 2015; Rogers et al., 2016; Van Lissa, 2022). Although, there are some examples. Self-regulation theory (Zimmerman, 2000) has been used to interpret, engineer, or organize features in the modeling process (Fan et al., 2021; Gašević et al., 2016; Matcha et al., 2019). Further, genetic algorithms have been used to incorporate theory into predictive modeling (Xing et al., 2015; Zhang et al., 2019). However, relatively few LA studies have integrated self-report survey data with validated measures of psychological constructs into their prediction models (Issah et al., 2023; Wong & Li, 2020). In a systematic review, Khanna et al. (2016) found that 69% of LA studies used prior academic achievement and demographic characteristics as the primary predictors. Recently, it has been hotly debated whether demographics should be included as predictors in relation to mitigating algorithm bias, improving prediction performance equally across demographic groups, and preventing overfitting (Baker et al., 2023; Cohausz et al., 2023; Deho et al., 2027; Yu et al., 2021).² To detect and understand potential bias, heterogeneity, and the importance of demographics in predictive models, model interpretability and the grounding of investigations in theory become even more relevant (Rogers et al., 2016).

There have been a growing number of emerging studies that have utilized machine learning methods to predict achievement outcomes from cognitive and noncognitive skills, with the goal to understand better the important predictors (Gamazo & Martínez-Abad, 2020; Kiray et al., 2015; Martínez Abad & Chaparro Caso López, 2017; Nadaf et al., 2021, 2022; Noetel et al., 2023; Psyridou et al., 2024; Yoo, 2018). For example, Yoo (2018) conducted an Elastic Net regression analysis to predict Korean fourth graders' math achievement scores with more than 150 predictor variables. The results suggest that some motivational variables (e.g., math self-confidence) have unique explanatory power above and beyond demographic variables. Martínez Abad and Chaparro Caso López (2017) used decision trees to predict general academic achievement in over 18,000 high school students in Mexico. They found that personal factors (particularly learning strategies, self-esteem, drug use, coexistence violence, and resources at home), followed by school-related and social factors, were the most important

predictors. However, all of these studies used cross-sectional data sets, providing little information on whether these factors predict change in math achievement and how the predictors change over a child's development.

Current Study

The current study addresses the following two research questions. First, we aimed to examine the extent to which math achievement scores can be jointly explained by the extensive set of predictor variables identified in the literature (i.e., cognitive ability, motivation and emotion, cognitive strategies, family context, classroom context, and demographic information) above and beyond prior math achievement scores. In other words, how do these predictors explain the change in math achievement relative to prior achievement? Second, we aimed to investigate how the predictive power of these predictors changes as students progress through the school years. To address the research questions, a machine learning methodology was chosen, given its advantage of being able to analyze a large number of possible predictors while also helping to protect against overfitting. Using machine learning, we analyzed the entire set of variables included in the longitudinal PALMA data set—project for the analysis of learning and achievement in mathematics (Pekrun et al., 2007). The PALMA study investigated adolescents' development in mathematics during secondary school in Germany. The present analysis focuses on data from Years 5 to 9. The data include an extensive set of variables (reported by students, parents, and teachers) that were considered to be relevant to mathematics competence by experts in educational psychology and mathematics education. Thus, the data is well-suited to address the research questions of the current study.

To examine developmental change in predictive relationships, we analyzed the data in two different ways. In the first analysis (called "next year prediction" analysis), we predicted math achievement scores at time $T + 1$ from all variables (including prior math achievement scores) at time T and examined how this relationship changed as students developed (i.e., we compare the results for predictors at different times T , where $T = \text{Year } 5$ up until Year 8). Thus, this analysis addresses the developmental change in the proximal predictive power of different variables in the data set. As previously noted, several studies have indicated that mathematics competence becomes more stable over time, and these findings suggest that the relative importance of other factors may decrease over time.

In the second analysis (called "varying lag prediction" analysis), we predicted future math achievement scores at different time intervals. Specifically, we compared how the initial assessment point (Year 5) variables predict math achievement scores at later time points (Years 6–9). Thus, this analysis addresses the change in the predictive power of variables assessed at the start of secondary school (Year 5) over different prediction intervals (i.e., assessing their temporal stability). Given that math achievement scores are not perfectly stable over time, we can naturally expect that the predictive power of the initial math achievement scores would decrease as the time intervals become larger. However, there has been little research comparing the predictive power of various antecedent variables (e.g., motivation and emotion variables) at different time intervals, and therefore, there were no specific expectations about these results.

² A subsequent review of EDM literature found that only 15% of studies included demographic features in their analyses (Paquette et al., 2020)

Method

Participants and Design

The data in this study were from the longitudinal PALMA study (Pekrun et al., 2007) (see also Arens et al., 2022; Frenzel et al., 2009, 2010; Marsh et al., 2022, 2019; Murayama et al., 2013, 2016; Pekrun et al., 2017, 2023, 2019), which included annual assessments of students in German secondary schools from Year 5 to Year 10. Samples were representative of the student population in the state of Bavaria (Germany) and included students from all three school types within the German public school system: vocational-track schools (Hauptschule), intermediate-track schools (Realschule), and academic-track schools (Gymnasium). In Bavaria, students typically enter these schools in Year 5 (i.e., the start of the PALMA study). Students in academic and intermediate-track schools tend to stay at least until Year 10, whereas compulsory vocational-track schooling ends after Year 9. Thus, we decided to focus only on data from Years 5 to 9 (see Murayama et al., 2013, 2016 for a similar approach).

Each year, students, teachers, and parents responded to survey questions toward the end of the school year. Students took an intelligence quotient (IQ) test and a standardized mathematics achievement test as part of the assessment. The Data Processing and Research Center of the International Association for the Evaluation of Educational Achievement conducted the sampling and the assessments. Trained external test administrators administered all assessments in the students' classrooms. Active parental consent was obtained for participating in the study. Participants were not provided any incentives. The studies of the PALMA project received Institutional Review Board approval from the Bavarian State Ministry for Education, Science, and the Arts (Reference: III/5-S4200/4-6/68 908).

In Year 5, there were 2,070 students from 42 schools (49.6% female, $M_{\text{age}} = 11.7$ years). Proportions of students in vocational, intermediate, and academic track schools were 37.2%, 27.1%, and 35.7%, respectively. In each subsequent year, the study tracked those who had participated in the previous assessment(s) and incorporated those who had not yet participated in the study but had become members of PALMA classrooms at the time of the assessment. As a result, the sample sizes from Year 6 to Year 9 changed in the following manner (Pekrun et al., 2007): 2,059 (50.0% female, $M_{\text{age}} = 12.7$ years), 2,397 (50.1% female, $M_{\text{age}} = 13.7$ years), 2,410 (50.5% female, $M_{\text{age}} = 14.8$ years), and 2,528 students (51.1% female, $M_{\text{age}} = 15.6$ years). Across all five assessments, a total of 3,425 students (50.0% female) and one of their parents, as well as the mathematics teachers of the participating classes ($N = 419$ teachers; 65.7% male) participated in the study. The actual sample sizes of students used in the present study for each analysis, as well as the % missing and % female, can be found in Table 1. Note that these are the final analysis data sets after preprocessing (i.e., dropping variables that had >50% missing data; see the Data Preprocessing section), and so the % missing constitutes the data that was later imputed as part of the machine learning pipeline (see the Data Analysis section).

Measures

The outcome variable of the current study is the standardized math achievement test score (see below). We then attempted to use all the remaining variables in the data set as predictor variables. Some

predictors were excluded, however, if they were either (a) not consistently assessed from Year 5 to Year 9 (except for the time-constant variables, e.g., sex), (b) had more than 50%³ missing data for any type of analysis we conducted, or (c) were superfluous (i.e., a recoding of an existing variable). When Pearson r correlations between variables were $>.7$, and it made conceptual sense to combine them, the mean score of the correlated variables was used to avoid unnecessary multicollinearity. All such aggregations are noted in the following subsections. In cases where it did not make conceptual sense to combine or remove variables based on collinearity (i.e., if by removing a variable, we would remove theoretically relevant information), variables were left in, and the possibility of multicollinearity affecting the interpretation of the model was dealt with at a later stage in the analysis (see the Model Interpretation section below). We had 88 predictors before dummy coding (see the Data Analysis section) and 105 predictor variables after.

The predictor variables were further classified into the following categories: (nonverbal) intelligence, motivation and emotion, cognitive strategies, student-rated classroom context, teacher-rated classroom context, family context, demographics and socioeconomic status (SES), and school track. We decided the groupings using a theoretical perspective (e.g., categories that are typically used in educational psychology) and the empirical data. Specifically, we aimed to ensure that Pearson r correlations between variables across different groups were not above .7 (this is important when interpreting the models; see the Model Interpretation section). Our approach represents a combination of “knowledge-driven” and “data-driven” strategies as defined in Au et al. (2022). We provide succinct explanations of the variables in each group below and a summary table in the online supplemental materials. When a construct was assessed using multiple items, we computed the average of the item scores (after accounting for reverse-coded items).

Math Achievement Scores

Mathematics achievement was measured using the PALMA Mathematical Achievement Test (Pekrun et al., 2007). This test comprises multiple-choice and open-ended items. It assesses students' modeling algorithmic competencies in arithmetics, algebra, and geometry. Importantly, the scores were scaled using Rasch modeling, which enabled longitudinal comparisons between different school years. The test had different versions for each school year. It was constructed using multimatrix sampling with a balanced incomplete block design (e.g., see PISA, 2018). Anchor items that were repeated across years were included to allow for the linkage across the five waves. To facilitate interpretation, we standardized all achievement scores ($M = 100$ and $SD = 15$) in relation to scores at Wave 1. Prior work has confirmed the test scores' unidimensionality and longitudinal measurement invariance (see, e.g., Murayama et al., 2013).

Motivation and Emotion Variables

The variables in this category included students' self-reports of noncognitive variables related to motivation and emotions, most of

³ The threshold of 50% was chosen so that the imputation model had at least the same amount of data to make predictions as predictions needed to be made, akin to a 50:50 train:test split in machine learning.

Table 1
Descriptives of Analysis Data Sets

Analysis	<i>N</i>	% Female	% Teacher missing	% Parent missing	% All missing
Next year prediction					
Year 5 → Year 6	1,822	50	6	5	3
Year 6 → Year 7	1,732	50	13	9	5
Year 7 → Year 8	2,168	51	31	14	9
Year 8 → Year 9	2,188	51	26	19	10
Varying lag prediction					
Year 5 → Year 6	1,822	50	6	5	3
Year 5 → Year 7	1,579	50	6	5	3
Year 5 → Year 8	1,471	50	6	5	3
Year 5 → Year 9	1,398	50	6	5	3

Note. We show the percentage of missing data across the teacher variables, parent variables, and all variables (student, parent, and teacher). Note that the percentages of missing data are the same (when rounded) for the varying lag analysis due to the data sets being subsets of the same data (Year 5) and dropout not affecting the distribution of missingness across variables.

which were contextualized around math. Specifically, these were competence beliefs in mathematics (16 items), achievement motivation (a total of eight items, comprising both performance-approach motivation and performance-avoidance motivation with four items each), intrinsic motivation (five items), interest (six items), and instrumental future-oriented motivation (three items). Furthermore, we assessed multiple mathematics-related values, including perceptions of the importance of math performance (five items), utility value (two items), intrinsic value (three items), and holistic value of mathematics (two items). Next, we assessed flow (six items) and task-irrelevant thinking tendencies (seven items), as well as effort (seven items). In addition, there were multiple scales measuring emotions associated with mathematics. For the present analysis, we grouped these scales as positive emotions (total of 17 items, comprising joy [nine items] and pride [eight items]) and negative emotions (total of 43 items, comprising anxiety [15 items], anger [eight items], shame [eight items], hopelessness [six items], and boredom [six items]). Lastly, we included one domain-general noncognitive variable, namely general self-esteem (six items).

Cognitive Strategies

This group comprised of student's self-reports on strategies used to learn math, including students' self-regulation of learning in mathematics (six items), external regulation of learning in math (six items), and different facets of elaboration, including procedural elaboration in terms of transfer of known task solution strategies (three items), declarative elaboration in terms of transfer of prior knowledge within mathematics (three items), and across other subjects or real life (three items). In addition, this group comprised mathematics study habits, including memorization of rules and task solution approaches (three items), rehearsal of rule application (three items), and creative problem solving (four items).

Student-Rated Classroom Context

The PALMA project measured both teacher and student ratings, with many constructs having parallel item wordings. As in prior studies (Murayama et al., 2013; Pekrun et al., 2023, 2007), teacher and student ratings of the same variables were not highly correlated (see Figure 2). Hence, we divided them into two groups. The first group included student reports on teacher behaviors and classroom experience. We

assessed the following student-perceived teaching behaviors: autonomy support in task solving in math (five items), support of self-regulated problem solving (five items), excessive versus adaptive pacing (two and four items), scaffolding and teaching to transfer (11 items), achievement pressure (six items), positive reinforcement (three items), punishment and support after failure (three and four items), variety in instruction (three items), and teacher enthusiasm (five items). In addition, students rated the degree to which mathematics class was disrupted (five items), time was wasted (four items), and students in the class had a positive attitude toward math (three items).

Teacher-Rated Classroom Context

This group contains teacher reports on classroom context and instructional behavior, teachers' collaboration with parents and colleagues, as well as information on the teacher's gender and job experience. Parallel to the students, the teachers rated autonomy support in task solving (five items), support of self-regulated problem solving (five items), adaptive pacing (two items), scaffolding and teaching to transfer (11 items), positive reinforcement (three items), punishment versus support after failure (three and eight items), variety in instruction (two items), and their own teaching and subject enthusiasm (eight items). Teachers also rated the degree to which mathematics class was disrupted (five items) and time was wasted (three items). In addition, the teachers rated their efforts in providing adaptive and comprehensible instruction (four items) and the degree to which they applied an individual (rather than social) frame of reference in judging students' work (four items). Finally, teachers reported on the quality of collaboration within the math department at their school (six items) and with parents (seven items).

Family Context

The PALMA project measured students' and their parents' ratings of the home environment, partly operationalized with parallel item wordings. When Pearson correlations across parallel-worded parent and student reports were larger than .70, they were combined using the mean. Generally, this group contains variables on parental attitudes, expectations, and activities in mathematics. Variables for which parallel-worded parent and student reports were combined included the expected and aspired math grade (one item each), the importance of the aspired grade (single items), parental instructional

support (six items each), and parental skill and enthusiasm for math (four items for student report and five items for parent report). Variables for which parallel-worded parent and student reports were used separately included parental autonomy support (five items each), family activities in math (six items parent report and five items student report), and positive attitude toward the math domain in the family (items each). Finally, there were four family context variables reported by students only, including parental achievement pressure in math (six items), parental positive reinforcement in math (two items), parental support after failure in math (three items), and general (nonmath-related) and joint cultural activities (four items).

Intelligence (Nonverbal)

Nonverbal intelligence was measured using the 25-item nonverbal reasoning subtest of the German adaptation of Thorndike's cognitive abilities test (Kognitiver Fähigkeitstest [KFT 4-12 + R]; Heller & Perleth, 2000).

Demographics and SES

This variable group included age, sex, early, or later entry into elementary school (i.e., entering school already at 5 years or later at 7 years, based on parents' and the elementary school's/kindergarten's joint decision), members of the household (categorical, as reported by both parents and students⁴), whether the student switched school track or not during the study, whether the students and their parents were born in Germany or not (as separate variables), whether the language spoken at home was German or not, and family SES. SES was assessed by parent report using the Erikson–Goldthorpe–Portocarero classification (Erikson et al., 1979), which consists of six ordered parental occupational status categories. We coded scores so that higher values represent higher family SES.

School Track

School track is a single categorical variable denoting which of the three school types within the German public school system the student belonged to vocational-track schools (Hauptschule), intermediate-track schools (Realschule), and academic-track schools (Gymnasium). Note that some students change track within the study time frame (typically initiated by teachers/schools). However, most changes from the vocational track to the intermediate-track from Year 6 to Year 7 were system-implied (i.e., defined by changes to the tracking system at the state level).

Data Preprocessing

We constructed different data sets required for our next year prediction and varying lag prediction analyses (see Figure 1), ensuring that variables were consistent across all data sets. We also ensured that the variables in each data set only enabled forward prediction in time. For example, in the next year prediction analysis, only grade information from 1 year before could be used as a predictor. We removed individuals who did not have data for the outcome variable for each data set. As described above, some aggregations of correlated variables were made when it made conceptual sense. These aggregations were made consistently across all data sets, even if the correlations exceeded the threshold in only one data set. Therefore, all data sets included the same variables. The

remainder of the preprocessing steps (imputation, categorical variable coding, and scaling) were performed within the analysis pipeline (see the Data Analysis section) to avoid information from the test data used to evaluate the models contaminating the data used to train the models (Lavelle-Hill et al., 2023).

Data Analysis

An overview of the analyses can be seen in Figure 1. We performed two different analyses, using either stable 1-year lags or varying time lags between the predictors and the outcome variable. For the first analysis, next year prediction, the time lag was always 1 year with a sliding window of 1 year (variables from Year 5 predicting the outcome at Year 6, Year 6 variables predicting the outcome at Year 7, etc.). This created four different data sets for the first analysis. In the second analysis, varying lag prediction, we used the Year 5 predictors to predict the outcome variable at Years 6, 7, 8, and 9 (lag = 1, 2, 3, and 4 years, respectively). For the second analysis, we created four further data sets. Note that there is one completely overlapping analysis in these two sets of analyses (predicting Year 6 achievement scores from Year 5 predictors). Thus, the results from these analyses are identical (but we display both for visual purposes).

The machine learning procedure was applied in a consistent manner across all eight analysis data sets. It consisted of the following steps: (a) splitting the data into a training set (80% of the full data) and a test (20%) set; (b) performing five-fold cross-validation⁵ on the training data to find the optimal model hyperparameters for two different machine learning models (Elastic Net model and Random Forest model, for more details see below); (c) predicting the 20% hold-out test data to evaluate the predictive performance on unseen data; and finally (d) interpreting the model using variable importance estimation methods to find the most predictive variables (or groups of variables). The machine learning pipeline was run separately for each data set. Here, we note that, like most common machine learning methods, we do not explicitly account for the hierarchical structure of the data (i.e., students nested within classes). Because the main reason we account for nested structures in traditional statistical approaches is the underestimation of sampling errors, and in the machine learning analysis, we do not use standard errors, we chose not to explicitly model the hierarchical data structure (for further discussion on this point, see Lavelle-Hill et al., 2023). To further support this decision, we note that the intraclass correlations of the outcome variable, after controlling for the school track and classroom context variables (which are predictors in our model), were low (Year 5: .08; Year 6: .05; Year 7: .09; Year 8: .09; and Year 9: .17).⁶

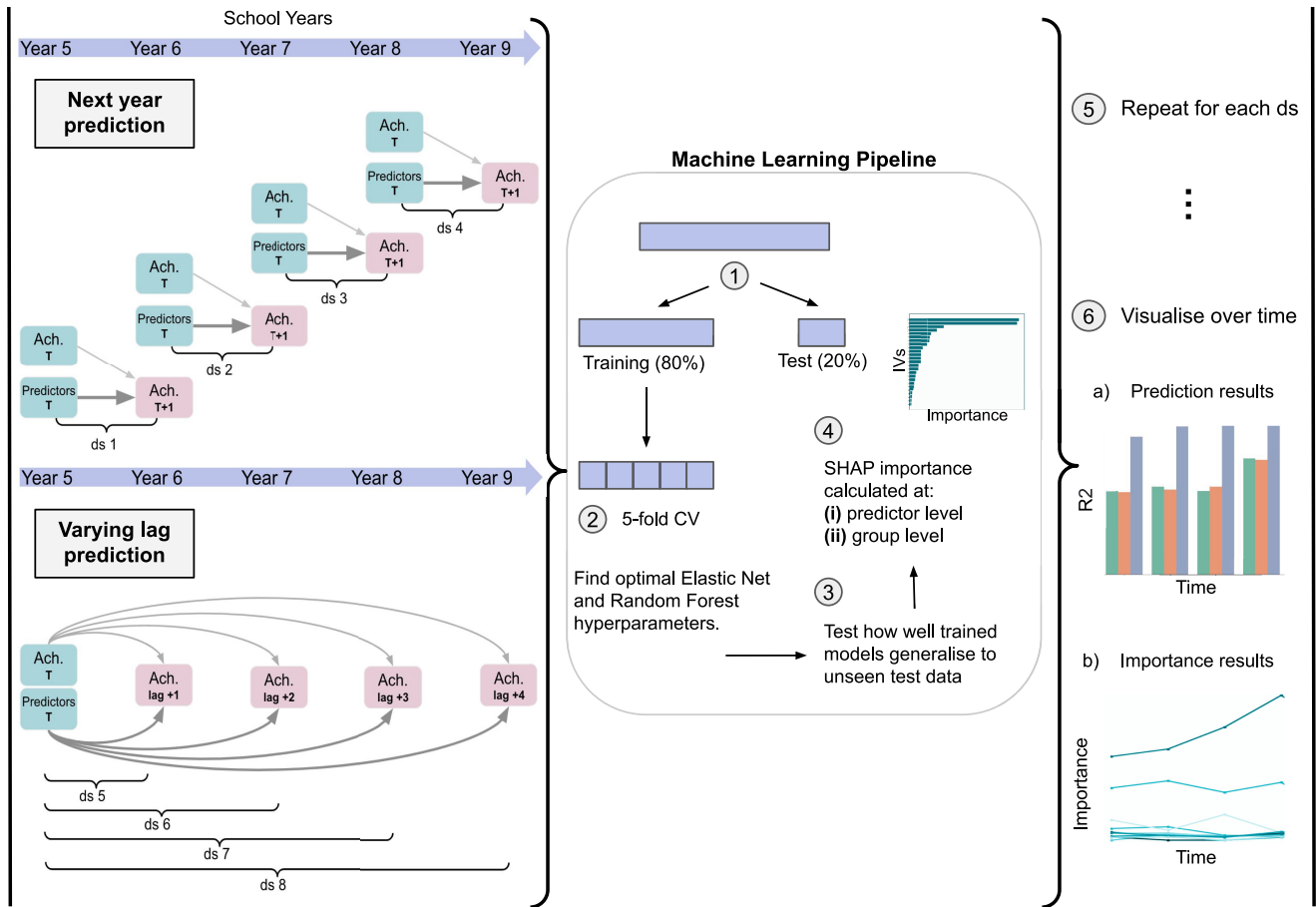
We chose two different machine learning models to predict our outcome variable: an Elastic Net model (Zou & Hastie, 2005) and a Random Forest model (Breiman, 2001). These models were chosen because they both allow the internal selection of only the most

⁴ There were some discrepancies between student and parent reports and thus both variables were retained in the model.

⁵ Five-fold cross-validation involves splitting the data randomly into five folds (or subsets). A model is then fit on four folds and tested on the remaining fold. This is then repeated five times so that each fold is the hold-out test data (often called "validation data") once.

⁶ These calculations were made using the residuals of a regression model using school track and classroom context variables to predict the outcome variable. Intraclass correlations were calculated using a mixed effects linear model, as there were unequal numbers of students in each class.

Figure 1
The Full Analysis Pipeline Predicting Achievement Scores in Mathematics (Ach.)



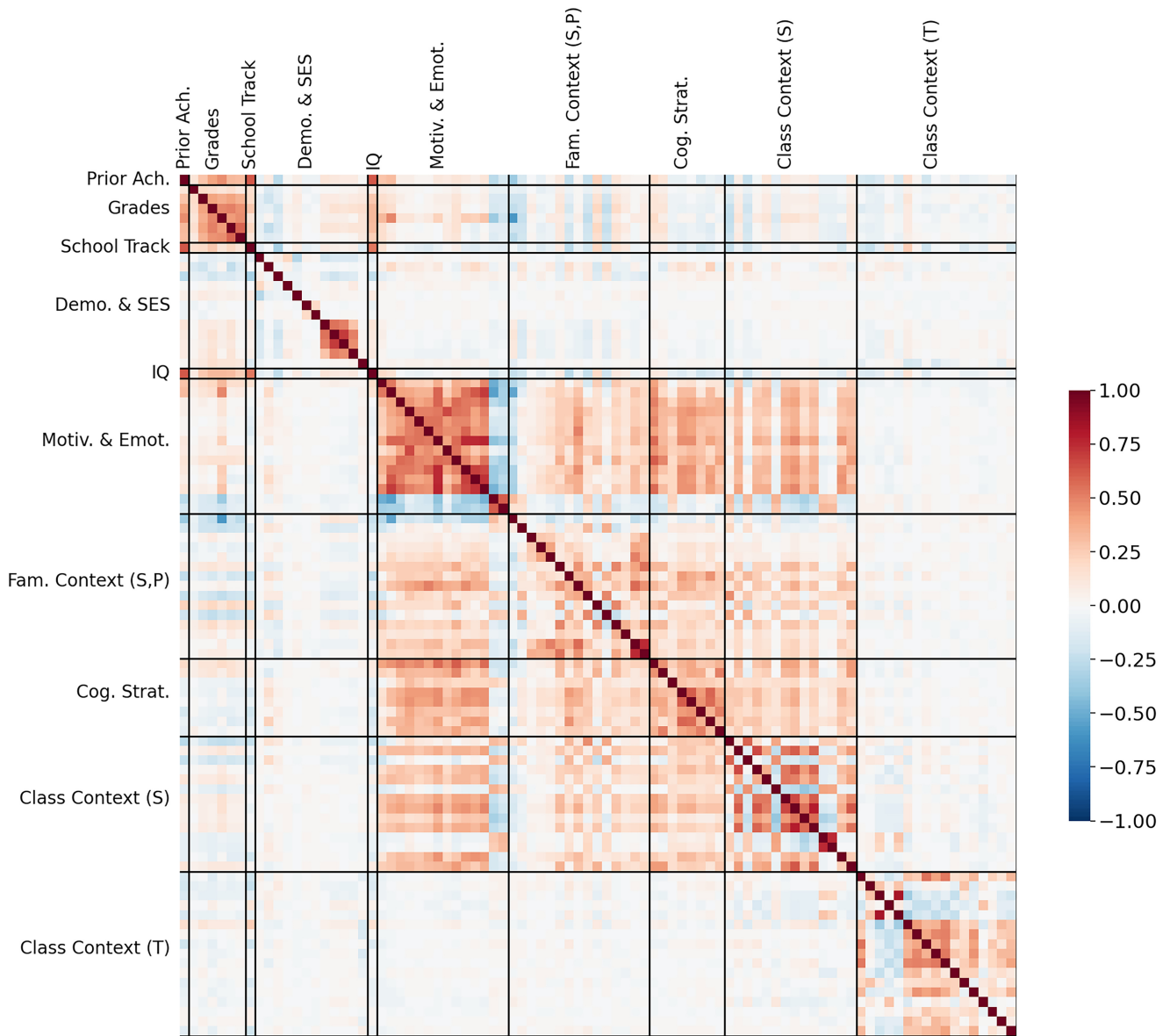
Note. Two core analyses were carried out: “next year prediction” predicting 1 year ahead using a sliding window of one year (i.e., from Year T to $T + 1$), and “varying lag prediction,” which used only the predictor variables (IVs) from Year 5 to predict the outcome at varying time lags into the future (e.g., from Year 5 to Year 9 is a lag of $+4$ years). Both analyses used the same machine learning pipeline to train and test predictive models for each data set (ds) and the same variable importance analysis for model interpretation. The prediction and variable importance results for each time point were presented together to identify temporal patterns. Ach. = achievement; T = measurement time for the predictors; $T + 1$ = measurement time for the predictors plus one year; ds = data set; IV = independent variable; CV = cross-validation; SHAP values = Shapley additive explanations values (Lundberg & Lee, 2017; see the Model Interpretation section). See the online article for the color version of this figure.

important variables from a large pool of possible variables. Thus, a priori variable selection or variable reduction methods are not needed. The Elastic Net model uses linear regression that combines the Lasso (L1 norm) and ridge (L2 norm) regularization penalties. Lasso encourages sparsity by imposing a penalty on small absolute magnitudes of regression coefficients (McNeish, 2015). This effectively performs variable selection, eliminating irrelevant or redundant predictors from the model. Ridge regularization, on the other hand, applies a penalty to the squared magnitudes of the regression coefficients. This helps to control for multicollinearity by shrinking the respective coefficients toward zero (Hastie, 2020). By combining both regularization techniques, the Elastic Net regression provides a flexible balance between variable selection and coefficient shrinkage. Therefore, the Elastic Net regression allows fine-tuning the strength of the regularization (α parameter) and a mixing parameter (the ratio of L1:L2 penalties) to define the type of regularization. The mixing parameter allows for a greater fine-tuning of the model based on the specific data

characteristics (Zou & Hastie, 2005). The grid of the hyperparameter options considered, as well as the optimal values selected, can be found in the online supplemental materials.

The Random Forest model comprises an ensemble of decision trees (Breiman, 2001), where each tree is built on a different random sample of the data and uses a different pool of possible predictors to select from at each split point (to maximize the variance between the trees). The predictions from all the trees are then combined through averaging or voting. Using binary decisions, each tree partitions the data into smaller and smaller subgroups (or “nodes”) to maximize the between-group variance and minimize the within-group variance. Thus, only the most optimal variables for splitting the data are used. Random Forest models are advantageous as they can automatically detect and model any nonlinearities, interactions, and subgroup effects within the data—without becoming too sculpted (or “overfitting”) to a particular sample (Breiman, 2001). The hyperparameters we tuned were the number of trees, the maximum depth of the trees, the minimum number of

Figure 2
Predictor Pearson r Correlations Sorted by Variable Groups



Note. Ach. = achievement; Demo = demographic; SES = socioeconomic status; IQ = intelligence quotient; Motiv. = motivation; Emot. = emotion; Fam. = family; S, P = both student- and parent-reported variables; Cog. Strat. = cognitive strategies; S = student-reported; T = teacher-reported. See the online article for the color version of this figure.

samples needed to make a split, and the number of predictors available for selection at each split point (see the online supplemental materials for the full grid and the final values that were selected).

Before training the models, some additional preprocessing steps were carried out as part of the machine learning pipeline. We applied these steps within each separate training and test (or “validation”) set to prevent possible information leakage between the data sets. These preprocessing steps included “dummy” (or “one-of-K”) encoding of categorical variables, data imputation, and data scaling. For the Elastic Net model, one category was always dropped as a reference class to eliminate the perfect dependencies between the dummy encoded

variables. This is not required for the Random Forest model, so all categories were retained in this analysis. We used two different prediction models to perform multivariate imputation to impute the missing data.⁷

⁷ Multivariate imputation was performed using the scikit-learn function `iterative_imputer` (Pedregosa et al., 2011a). The function was inspired by the multivariate imputation by chained equations (MICE) package available in R (Van Buuren & Groothuis-Oudshoorn, 2011), but only returns a single imputed value instead of multiple imputations (this is common in machine learning methodology when standard errors are not being estimated—for more discussion on this, see Lavelle-Hill et al., 2023).

We used a ridge regression for numerical variables, and for categorical variables, a Random Forest classifier.⁸ A multivariate imputation method iterates over the columns in a round-robin fashion, predicting the missing values in the target column using the other variables in the data (Pedregosa et al., 2011b). Note that no missing values from the math achievement scores were imputed. This is because, by imputing the missing outcome variable, machine learning models can learn the function used to impute the missing data (rather than the true relationships between the predictors and outcome for the data that is present) (Lavelle-Hill et al., 2023). So, if a student did not have outcome information, they were dropped from our sample. Data were then centered and scaled using *Z*-score transformations (where the mean is subtracted from each data point, and the result is divided by the standard deviation).

Model Interpretation

To understand which variables (or groups of variables) are important predictors and how these change over time, we computed variable importance estimates using Shapley additive explanations (SHAP) values (Lundberg & Lee, 2017). SHAP is a post hoc model-agnostic method for interpreting prediction models. SHAP is based on Shapley values from cooperative game theory, used to assign payouts to players depending on their contribution (Shapley, 1997). SHAP calculates the contribution of each variable to the overall prediction. This is done by calculating the marginal average contribution of the variable across all possible (ordered) combinations of variables in a model. A SHAP value is initially calculated for each instance (in our case, an instance is an individual) in the data. These instance-wise SHAP values are directional (i.e., a positive value indicates a positive relationship with the predicted outcome). The absolute SHAP values can also be summed over all instances in the data set to provide an overall measure of importance for each variable in the model. This overall measure of predictor importance is the primary focus of the present analysis, although instance-wise SHAP values are also subsequently presented to analyze heterogeneity in the effects (see below).

Although there are alternative model-agnostic methods to compute predictor importance, such as permutation importance (Altmann et al., 2010) or local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016), we argue that SHAP values have three key benefits: (a) They indicate the direction in which each predictor contributes to the prediction (i.e., whether there is a positive or negative predictive relationship); (b) SHAP values are calculated first on the level of each data instance (i.e., here, each individual student), enabling an analysis of whether the variable has the same predictive effect for all individuals or whether there is heterogeneity (e.g., for some individuals, the variable pushes the prediction up, and for others down); (c) they are additive, and can easily be aggregated to produce an explanation at the level of the whole sample.

To facilitate the interpretation of the results (and further reduce the impact of multicollinearity), we decided to analyze the importance of individual predictors and groups of similar variables. We computed the absolute SHAP importance for each group of variables as defined in the Measurement section (i.e., motivation and emotion variables, cognitive strategies, student-rated classroom context, etc.). This way, we derive a set of interpretable importance values for each group of predictors. The group importance was calculated by first calculating the importance of each predictor in the model across all individuals and then summing up the absolute SHAP values for all variables within

the group to get a combined group SHAP importance score. These more stable group variable importance scores are the main focus of our results.

Transparency and Openness

All code required to rerun the data preprocessing as well as the analyses can be found on GitHub: <https://github.com/Rosa-Lavelle-Hill/palma-ml-open>. The raw data and the survey questions (in German) can be made available upon request.

Results

Before running the machine learning analysis, we first checked whether our variable aggregation and grouping worked empirically with regard to the issue of multicollinearity. Figure 2 provides a heatmap of the correlations between predictors (averaged across data sets). As can be seen, the correlations between groups are generally smaller than the correlations within groups, supporting our process of grouping predictor variables using conceptual similarity and Pearson *r* correlation values.

Next Year Prediction

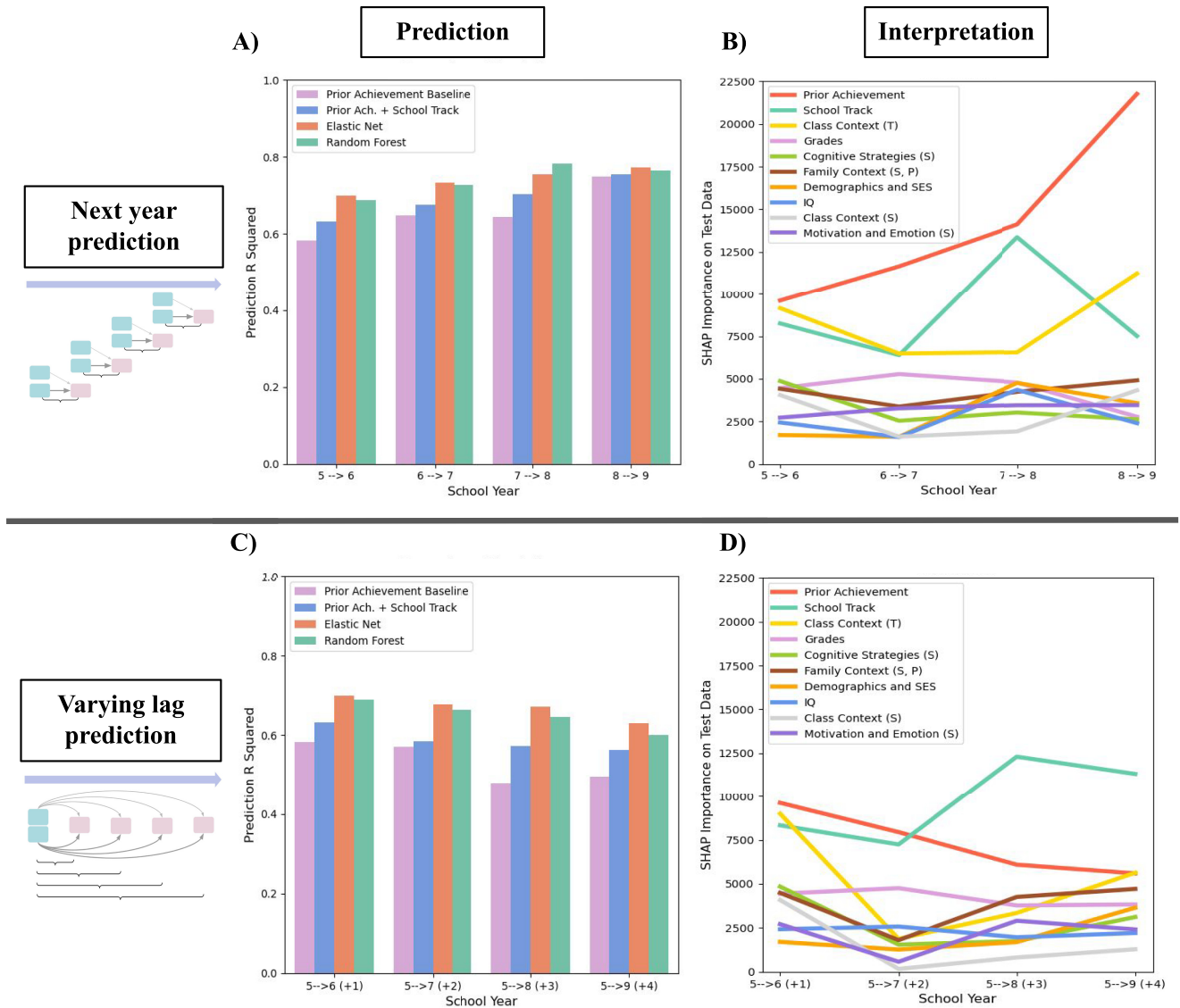
The first analysis, next year prediction, used all predictor variables, including prior math achievement, to predict math achievement 1 year ahead. The predictive performance of two models (an Elastic Net model and a Random Forest model) can be found in Figure 3. Panel A shows the model performance using prediction *R*-squared values. This is similar to *R*-squared values in standard regression, but it quantifies the degree to which the machine learning model can predict math achievement scores for new observations (Scheinost et al., 2019). Note that a prediction *R*-squared value of 0 is equivalent to the performance of a naive model using only the mean of the outcome variable in the training data to predict the outcome in the test data. We found that the best machine learning model could predict future achievement scores well, with the prediction *R*-squared values: Year 5 → 6 = .70 (mean absolute error [MAE]⁹ = 34.99), Year 6 → 7 = .73 (MAE = 34.44), Year 7 → 8 = .78 (MAE = 32.74), and Year 8 → 9 = .77 (MAE = 35.97). Note that the MAE should be interpreted in relation to baseline measures; for the MAE plotted next to the mean and prior performance baselines, see the online supplemental materials.

The machine learning models' prediction *R*-squared values are compared to two baseline models in Figure 3A. The first baseline is the prior achievement-only model, which predicts math achievement scores only using prior achievement (on the same standardized math test). The second baseline additionally includes information on the school track. Both baseline models use ordinary linear regression analysis. By comparing the full model with the baseline models, we identified the extent to which the other predictor variables collectively predicted the *T* + 1 math achievement scores (the outcome variable) above and beyond the prior math achievement scores (at *T*). As can be seen from the pink bar in Figure 3A, prior math

⁸ This method is akin to the `missForest` package in R (Stekhoven & Bühlmann, 2012).

⁹ Note that the MAE is simply the absolute mean of the residual error (calculated as actual values in the test data minus the values predicted by the model).

Figure 3
The Prediction and Variable Importance Results



Note. (A, B) The analysis predicting 1 year ahead (next year prediction) with a sliding window of 1 year; (C, D) The analysis with a varying time lag (varying lag prediction). (A) and (C) The prediction performance of the Elastic Net model and Random Forest model to two baseline models: only the prior math achievement scores at the time the predictor variables were measured (without any other questionnaire data); and prior math achievement combined with school track. (B) and (D) The interpretation of the best Elastic Net model at each time point using SHAP variable importance values aggregated to the group level. Ach. = achievement; SHAP = Shapley additive explanations; T = teacher-reported; S = student-reported; S, P = both student and parent-reported variables; SES = socioeconomic status; IQ = intelligence quotient. See the online article for the color version of this figure.

achievement scores predict future achievement scores with a prediction *R*-squared of .58 (Year 5 → Year 6), .65 (Year 6 → Year 7), .65 (Year 7 → Year 8), and .75 (Year 8 → Year 9), respectively. The increasing trend suggests that prior math achievement has increasing predictive power as students progress through secondary school—this finding is further confirmed in the variable importance analysis (see below). School track has some added prediction power (blue bar in Figure 3A). Importantly, the other predictor variables also seem to have added predictive value. These additional effects appear to be the smallest in the model predicting the final year outcome. More

specifically, by comparing the prior performance and school track baseline model to the best-performing machine learning model, self-report data, grades, and cognitive tests add an additional prediction *R*-squared of .07 (Year 5 → Year 6), .06 (Year 6 → Year 7), .08 (Year 7 → Year 8), and .02 (Year 8 → Year 9), respectively.

We computed group SHAP values to analyze the importance of groups of predictors of math achievement and how they change over time. We did this for both the Elastic Net and Random Forest models. For simplicity, we present just the Elastic Net model results here because it is the simplest and best-performing model for three

out of four time points. The variable importance results are illustrated in Figure 3B. There are a few notable observations. First, prior math achievement appears to be one of the strongest predictors of math achievement across all intervals. Second, consistent with Figure 3A, the influence of prior math achievement indeed becomes stronger over time, indicating that the interindividual stability of achievement scores increases as students progress through the school years. Third, aside from the school track being highly important for Year 7 predicting Year 8, classroom context rated by teachers is the strongest predictor among the other predictor variables and takes on a “U”-shaped trend (being most predictive at the start and end of secondary school). Finally, while other factors (motivation and emotion, cognitive strategies, and family context) have relatively small contributions, the effects are relatively temporally stable, despite the importance of prior achievement becoming stronger over time.¹⁰

We also analyzed the effect of each variable at the individual level by plotting the SHAP values of individual students in Figure 4 (for more details on how these are calculated, see Lundberg & Lee, 2017). For this analysis, we used the Random Forest model, which enabled us to identify any potential nonlinear relationships. It is important to highlight here that the Random Forest model is not the same as the model depicted in Figure 3B (Elastic Net regression). Due to the different underlying model mechanisms, the two models (Elastic Net model and Random Forest model) will use the features differently, thus likely producing slightly different predictions and variable importance rankings. Therefore, an exact mapping of the results cannot be made between the two figures.

In Figure 4, each panel is a different prediction year, moving forward in time from top to bottom. The x axis represents the grouped predictors. For each predictor, a dot represents an individual student in the data. The dot’s color represents the predictor’s value for that individual (red = high positive; blue = high negative, as data were centered on zero). SHAP values are depicted along the y axis, which is essentially the gain or loss (compared to the overall mean) in predicted math achievement for that individual when that predictor variable is in the model. Note that the SHAP value scores plotted in Figure 3B and 3D are simply the sum of the absolute SHAP values from all the individuals and the variables in the group—but for the Elastic Net model, not the Random Forest model.

There are two important pieces of information we can derive from the plot. First, if many of the dots are highly positive or negative on the y axis, the predictor makes a strong contribution to the outcome variable (either positively or negatively). This means that higher or lower achievement scores were predicted for individuals as a result of the value of this predictor. Second, if the color of the dots for a certain predictor does not change smoothly from high (positive) values to low (negative) values (i.e., red [dark gray] to yellow [light gray] to blue [dark gray]) or vice versa, this is an indication of potential nonlinear effects. This is because the change of the predictor variable would not be linearly associated with the gain or loss in the math achievement score.

We observe that, in general, the effects appear linear and homogeneous, and there is no strong indication of nonlinear main effects, which is likely why the Elastic Net model (i.e., the linear machine learning model) performed so well. Nevertheless, Figure 3A shows that the Random Forest model was superior to the Elastic Net model for Year 7 → Year 8. However, using Figure 4, we cannot immediately identify a nonlinear main effect. Therefore, it is likely that the superiority of the Random Forest model for Year 7 →

Year 8 comes from other sources of nonlinearity that are not depicted in this graph, such as the modeling of interaction effects.

As Figure 4 displays the importance values for individual features, we can see that the single classroom context variables on their own do not appear to be highly predictive. Instead, the importance is distributed fairly evenly among the variables in the group, and it is their combined effect that matters. This is in contrast to IQ, which is measured using multiple items but recorded as one variable in our data. Figure 4 shows that, in the Random Forest model, IQ as a single variable is more predictive than any of the class context variables on their own.

Another observation from Figure 4 is that the individual variable importance measure that varies the most between the years is students’ math grades. In Year 5 → Year 6, math grade is an important linear predictor, where a higher prior grade strongly predicts higher math achievement. This relationship becomes less clear as time goes on, and by the final prediction (Year 8 → Year 9), the effects are not obvious (in fact, grade in music, the variable to the left, is more important). It is likely that as time passes and prior math achievement becomes more important, the predictive capacity of math grade will already be captured by prior achievement. In other words, in the later years, there is less additional predictive utility in math grades when prior achievement is in the model.

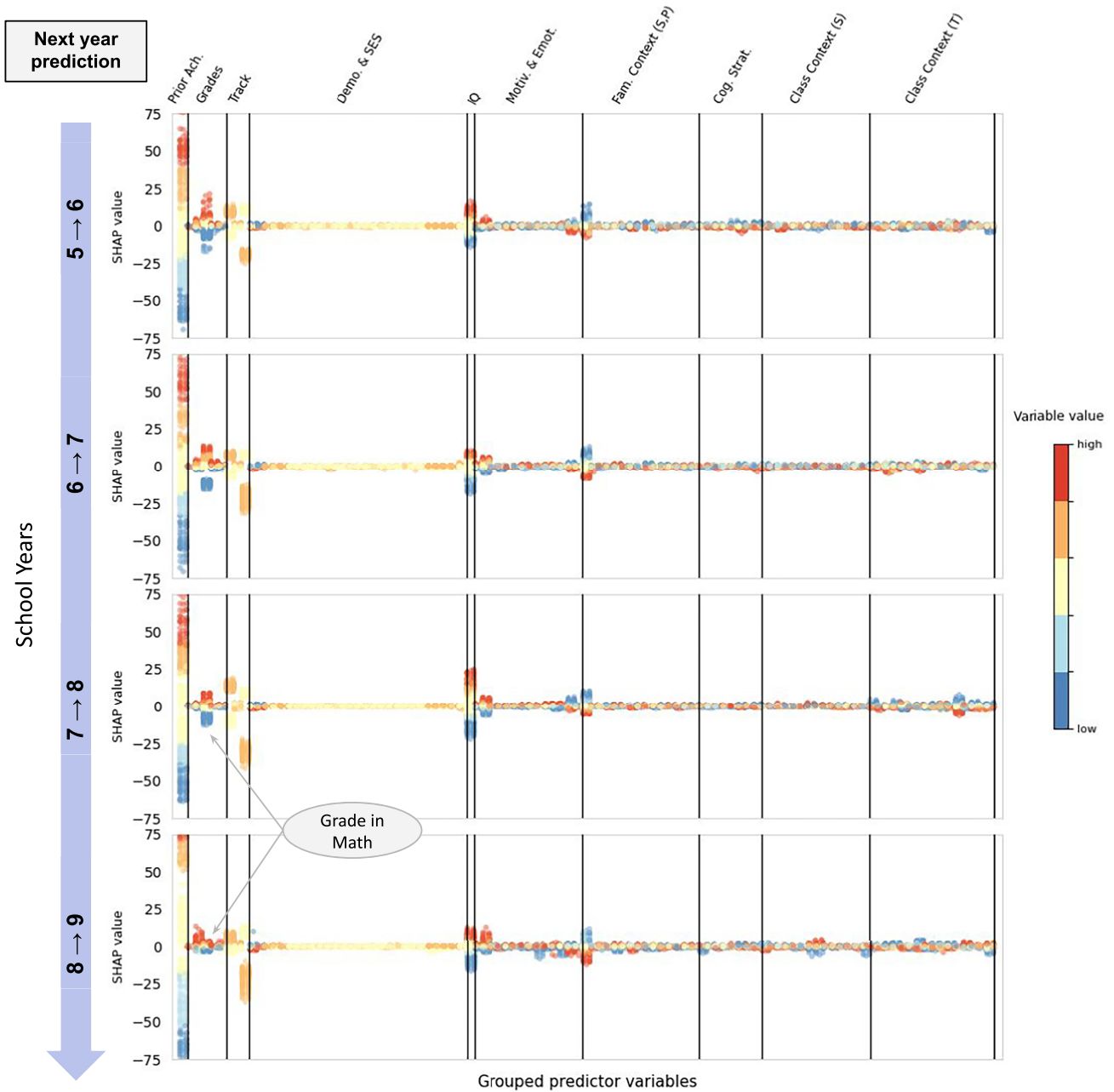
Varying Lag Prediction

To better understand the different variables’ predictive capabilities when predicting further into the future, we varied the number of years ahead of time that math achievement was predicted. For this analysis, the same set of predictor variables from Year 5 were used to predict math achievement at Years 6, 7, 8, and 9. The prediction performance can be found in Figure 3C. The best machine learning model could predict future achievement scores with prediction R -squared values of Year 5 → Year 6 = .70 (MAE = 34.99), Year 5 → Year 7 = .68 (MAE = 34.52), Year 5 → Year 8 = .67 (MAE = 39.48), and Year 5 → Year 9 = .63 (MAE = 43.16). Note, again, that the error should be interpreted in relation to baseline measures; for the MAE plotted next to the mean and prior performance baselines, see the online supplemental materials. As can be seen by the pink bar in Figure 3C, the predictive power of prior achievement (i.e., achievement at Year 5) decreases as the interval between the predictors and outcome becomes longer, with a particular drop for Year 5 → Year 8 (lag = 4). The prediction R -squared values for the prior achievement baseline are .58, .57, .48, and .50, respectively. These findings indicate that it is more difficult to predict future math achievement scores solely from the Year 5 achievement when the time interval is larger, specifically 3 years or longer. The analysis also showed that the school track added additional predictive power, which somewhat increased in later years. Finally, the analysis demonstrates the additional predictive utility of other predictors (i.e., survey data, grades, and IQ) was greatest at the longer time lags. Interestingly, the additional

¹⁰Note that there is not a direct relationship between the prediction R -squared values in Figure 3A and the SHAP importance in Figure 3B because (i) in Figure 3A, the model performance includes not just the sum of predictive utility of the individual variables, but also their interactions; and (ii) SHAP importance is just one way of measuring importance, and is not directly related to the final prediction R -squared of the model.

Figure 4

Graphs Showing the Directional and Individual Level Effect of the Variables on the Prediction of Math Achievement for the “Next Year Prediction” Analysis



Note. Here, the Random Forest model is used to enable a visual analysis of potential nonlinear effects. Each dot represents a different individual. The y axis depicts the directional SHAP value so that a positive SHAP value represents a predicted increase in math achievement, and a negative value indicates a predicted decrease. The color bar indicates the predictor variable’s value, where red (dark gray) is a high positive value, yellow (light gray) is a value close to zero, and blue (dark gray) is a high negative value. The (school) track and the demographic and SES predictors are binary with values 1 (orange [medium gray]) and 0 (yellow [light gray]); all other variables’ values are normalized and centered on 0. Ach. = achievement; Demo = demographic; SES = socioeconomic status; IQ = intelligence quotient; Motiv. = motivation; Emot. = emotion; Fam. = family; S, P = both student and parent-reported variables; Cog. Strat. = cognitive strategies; S = student-reported; T = teacher-reported; SHAP = Shapley additive explanations. See the online article for the color version of this figure.

predictive benefit of the other variables, beyond prior achievement and school track, seems to be relatively stable regardless of the time lag: .07 (lag = 1), .09 (lag = 2), .10 (lag = 3), and .07 (lag = 4).

The results from the variable importance analysis in Figure 3D confirm the aforementioned observations. First, prior math achievement scores become less important when predicting math achievement

further into the future. Instead, school track in Year 5 is an important predictor of math achievement, particularly for predicting Year 8 and Year 9 math achievement (lags = 3–4). Classroom context in Year 5 (as reported by the teacher) is also important, although less so when predicting Year 7 achievement. Classroom context (as measured by both the teachers and students) shows a drop in importance when predicting Year 7 achievement, the year where the students in the sample had the greatest changes in teachers. Notably, groups of variables measured using self-report questionnaires and cognitive tests (motivation and emotion, IQ, family context, and cognitive strategies) continue being important in predicting math achievement up to 4 years into the future, even when the overall predictive power of the model drops off.

Figure 5 allows us to look for any nonlinear effects or changes in how the model uses the individual predictors over time. Each panel represents a different year's math achievement being predicted, with the time lag increasing from top to bottom. As with the next year prediction analysis, the effects appear mostly linear. However, there is some heterogeneity across individuals, represented as "outlier" dots. Specifically, there are some individuals for whom grades were more predictive than their peers. A second observation related to the bottom panel (Year 5 → Year 9, lag = 4) is that the classroom context variables (as reported by the teacher) are more predictive relative to other time lags. This can also be seen in Figure 3D, where the classroom context (T) group, denoted by the yellow line, is highly important when predicting four years into the future. Teachers' enthusiasm is particularly important, as indicated in Figure 5. Therefore, we can infer that the higher importance of classroom context when predicting achievement 4 years into the future is largely attributable to the teacher's enthusiasm in Grade 5.

General Discussion

This study used a machine learning approach to predict and better understand the development of math achievement in secondary school using longitudinal survey data. Overall, we find that math achievement can be predicted to a high degree of accuracy, even up to 4 years into the future. Importantly, we gained several interesting observations by examining the predictive utility of different groups of predictors, which we discuss below.

Substantive Implications

When predicting the subsequent year's math achievement, the results show that as students go through the school years, prior math achievement becomes more important as a predictor. This indicates that the interindividual stability of achievement scores increases as students progress through secondary school. Previous study findings have aligned with this trend (Geary et al., 2017; K. Lee & Bull, 2016; Lin & Powell, 2022). Our results confirm this with a large representative sample of longitudinal panel data. This trend is also consistent with common findings that the genetic influence on achievement scores increases across school years (Selzam et al., 2017; von Stumm et al., 2020). In fact, increased genetic effects would be expected to lead to increased stability. Genetic effects can increase stability because genetics are stable and can exert stable effects over time, while environmental effects are unstable and can lead to increased variability. Importantly, with the increased stability of achievement scores, the relative contribution of other predictors

(e.g., cognitive, motivational, affective, and parental factors) decreases (e.g., Figure 3A). These findings suggest that educational interventions (i.e., remedial tutoring) may be most effective when provided early on when the link between existing achievement and future achievement is still relatively weak. Thus, in relation to the question of whether there is a "critical period" when interventions should be conducted to improve students' math achievement, our results suggest that the earlier in secondary school, the better.

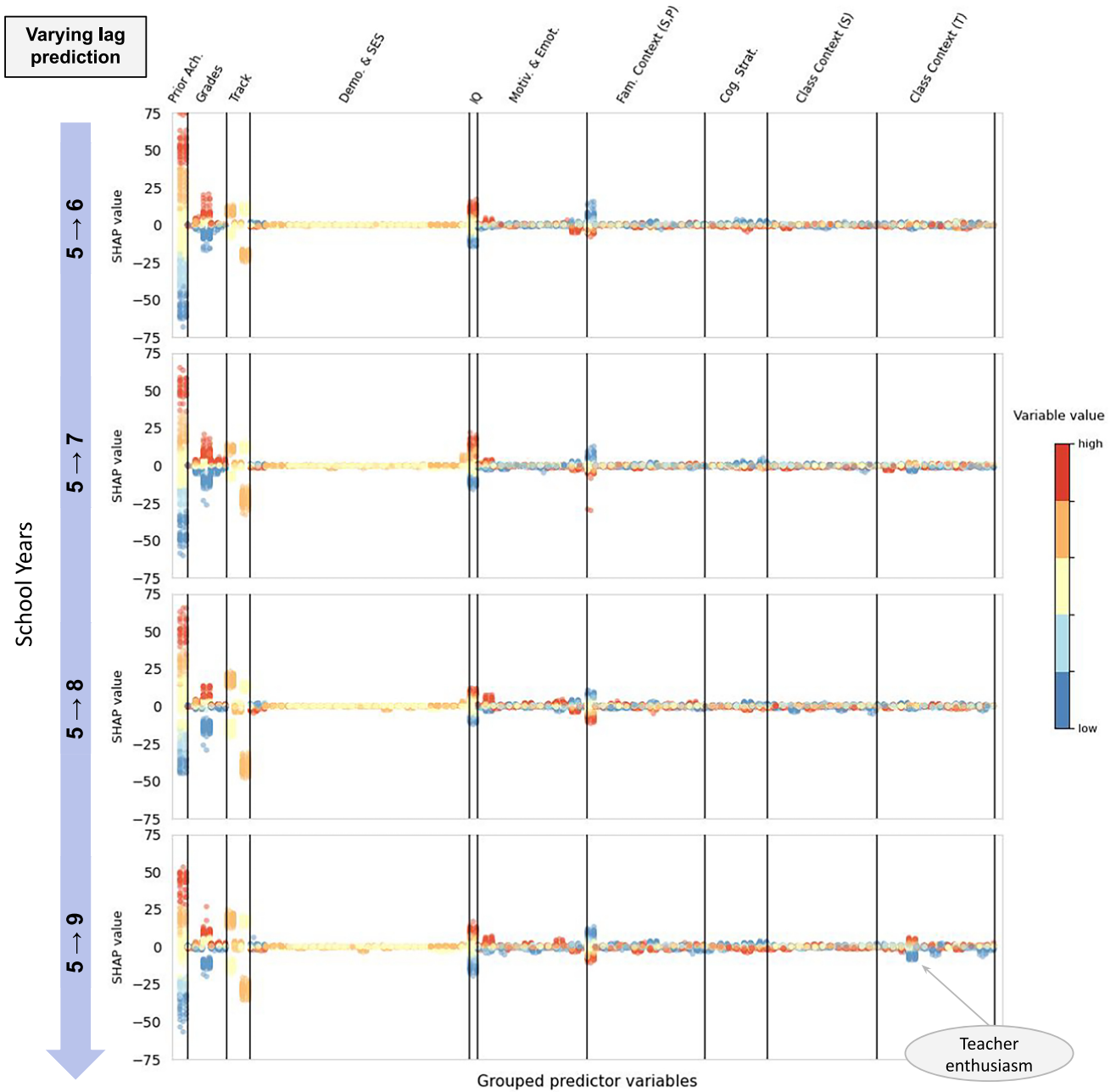
In the next year prediction analysis, the influence of school track was strongest where Year 7 was predicting Year 8. At the time of the PALMA study, a policy change was being made in Bavaria. The old policy implied that by the end of Year 4, the high-performing students would be sent directly to Gymnasium, while all other students were educated in a comprehensive system, and the differentiation between intermediate-track and vocational-track only happened after Year 6. The new policy implied that students would be streamed into either intermediate or occupational track directly after Year 4. The PALMA sample comprised students from both policies. In regions where the old policy was still in place, new intermediate schools were recruited after Year 6 if they took on a considerable number of PALMA cohort students. As a result, many students in the PALMA sample changed track between Year 6 and Year 7 (from comprehensive—coded as vocational—to intermediate). This could explain why the school track became more predictive when predicting next year's math achievement after this reshuffle period. However, as we cannot know which track changes resulted from the policy change and which were initiated by the school/teacher, it is difficult to hypothesize further about these results.

Second, the next year prediction model also shows that teacher-reported classroom context is the strongest group of predictors of math achievement after prior math achievement and school track. This is an important finding, as it is also a highly changeable predictor, which could be manipulated in practice (compared to more stable variables such as prior achievement). Classroom context includes various instructional strategies that teachers use during a class. The importance of teaching and instructional quality has been previously documented in educational research (e.g., Kunter et al., 2013), and our results corroborate these findings. Interestingly, the classroom context variables were more important than the home environment, motivation, and emotion variables. It is worth noting, however, that these findings emerged in the present study, which had mathematics as the focal domain. Formal learning at school is clearly the predominant driver of secondary-level mathematics skill development. In contrast, home environments are less likely to provide informal learning opportunities for those skills—this might be different for other domains, such as the literature or the arts. The effects of classroom context on math achievement were strongest at the beginning and end of secondary school (Year 5 → Year 6 and Year 8 → Year 9). In Bavaria, at the time of the data collection, Year 7 is when students had their largest changes in teachers (i.e., students were more likely to have the same teachers in Years 7 through to 9, which would be different from those they had for Years 5 and 6). This could explain the "U"-shape in the importance of the classroom context variables across secondary school years.

Moreover, the teacher-reported classroom context was much more predictive than students' reports about the classroom, even though most of the measures used the same items for teacher and student reports. Furthermore, student reports were not highly correlated with

Figure 5

Graphs Showing the Directional and Individual Level Effect of the Variables on the Prediction of Math Achievement for the “Varying Lag Prediction” Analysis



Note. Here, the Random Forest model is used to enable a visual analysis of potential nonlinear effects. Each dot represents a different individual. The y axis depicts the directional SHAP value so that a positive SHAP value represents a predicted increase in math achievement, and a negative value indicates a predicted decrease. The color bar indicates the predictor variable’s value, where red (dark gray) is a high positive value, yellow (light gray) is a value close to zero, and blue (dark gray) is a high negative value. The (school) track and the demographic and SES predictors are binary with values 1 (orange [medium gray]) and 0 (yellow [light gray]); all other variables’ values are normalized and centered on 0. Ach. = achievement; Demo = demographic; SES = socioeconomic status; IQ = intelligence quotient; Motiv. = motivation; Emot. = emotion; Fam. = family; S, P = both student and parent-reported variables; Cog. Strat. = cognitive strategies; S = student-reported; T = teacher-reported; SHAP = Shapley additive explanations. See the online article for the color version of this figure.

teacher reports (see Figure 2). This suggests that useful and important information about teaching and classroom environment can be collected from teachers rather than students. However, one possible

methodological reason why teacher-reported variables were especially predictive could be that teacher reports—unlike student reports—were not highly correlated with other student-reported predictors in the

model, allowing them to explain additional variance in the achievement scores. For example, Figure 2 shows student-reported classroom instruction variables were correlated with students' motivation, emotions, and use of learning strategies (see also Jaekel et al., 2021). In order to rigorously test the utility of teacher versus student reports, however, the teacher-reported variables should be pitted against the student-reported variables aggregated at the classroom level. Overall, the findings suggest that teacher-reported information has a unique predictive value and highlight the advantages of gaining information from different stakeholders' perspectives.

For the varying lag prediction analysis (i.e., Year 5 math achievement predicting future math achievement with varying time lags, from 1 to 4 years), not surprisingly, the further into the future we predicted, the less important Year 5 math achievement became. This simply means that future prediction of math achievement from achievement in Year 5 becomes more difficult as the prediction interval increases. Instead, the school track students were in at Year 5 became more important as the time lag increased. Importantly, however, other groups of predictors, such as motivational, emotional, cognitive, classroom, and family factors, remained stable in their predictive importance regardless of the time lag, up to 4 years into the future. As a result, the effects of these predictors became more important relative to prior achievement. Therefore, it is likely that these processes are slower to take effect but are still critically important long-term, or when the effect of prior achievement is reduced. This is consistent with many findings in psychology that have many smaller cumulative effects (Götz et al., 2022) as well as theoretical perspectives indicating that motivational and socio-emotional factors can have long-lasting effects (e.g., Murayama, 2022; Yeager & Walton, 2011). However, while some studies suggest the importance of these factors for long-term rather than short-term prediction (e.g., Murayama et al., 2013), to our knowledge, little research has systematically examined the stability of the effects of these factors on math achievement. In fact, many of the statistical models that researchers typically use in longitudinal data analysis (e.g., lag-1 cross-lagged panel models) implicitly assume that the effects of these factors would decrease over time. Future studies may want to investigate the relative longevity of the effects of various student and context factors on math achievement.

Finally, although the main goal of this article was to produce theoretical insights to be used in research in educational psychology, we highlight that such insights could also be useful for practitioners and other stakeholders who might design interventions. For example, we find that the relationship between past and future performance increases as students progress through the school years, suggesting that educational interventions are likely most efficient early on. However, for several reasons, we do not envisage our predictive model being directly used as a tool by schools in practice. First, most schools outside of the context of this study would not have access to the vast amounts of questionnaire data used to train our model. Second, even within the sample of schools in this study, data would ideally need to be collected and updated annually, which would be extremely resource-intensive. Our results do, however, suggest that resources might be best directed toward collecting data from teachers about the classroom context. Finally, there are additional ethical questions that are raised when models are applied in practice, particularly in relation to the use of demographic variables and potential biases present in the data (Baker et al., 2023; Cohausz et al., 2023; Deho et al., 2027; Yu et al., 2021).

Interestingly, our model does not appear to overly rely on demographic features to make its predictions. This could be due to prior achievement, school track, and the classroom context variables being able to explain any differences between demographic groups. However, if our model were to be deployed in practice, its performance would need to be evaluated across different demographic groups to check for fairness and possible bias.

Methodological Implications

In most cases, we found that the linear model (the Elastic Net model) was the better-performing machine learning model. This, alongside analyzing the patterns of SHAP importance in Figures 4 and 5, suggests that the relationships between the predictors in the data set and math achievement are linear and independent—the impact of nonlinear or interaction effects appeared limited. Similar results have been observed before for survey data using machine learning methods (e.g., Jacobucci & Grimm, 2020; Lavelle-Hill et al., 2020; Salganik et al., 2020). The detection of interaction effects in survey data is often underpowered (McClelland & Judd, 1993), which may be why we did not observe strong interaction effects. However, it is important to note that we only compared one linear and one nonlinear model in our analysis. Thus, a different nonlinear model (such as an XGBoost algorithm; Chen & Guestrin, 2016) may have performed better than the Random Forest algorithm we used. Despite this, we decided against training more complex models after we noticed the superiority of the simpler linear model—in line with the recent discourse of “simple is better” when producing explanations from machine learning models (Rudin, 2019).

Regardless of the reason for the superiority of the linear model in our analysis, the present results imply that the utility of a machine learning approach in this kind of data set may be less in modeling nonlinear effects and interactions but in being able to consider many different predictors at once, while also guarding against overfitting. Importantly, we combined the data-driven approach of machine learning methods with theory-driven survey design (using reliable and valid measures of psychological constructs), data collection, and variable grouping (plus a rigorous methodology for interpretation). We believe that this type of hybrid approach could be an important pathway forward for applying machine learning methods to educational data, helping to answer the call for greater integration of theory with predictive modeling in educational research (Rogers et al., 2016).

When interpreting our findings, it is important to highlight that our predictive models are constructed in a data-driven manner rather than top-down using causal inference, and therefore, like with other regression models, we should not interpret the findings as causal. Although variables were only input into the model if there was some theoretical reasoning about their possible causal relationship, with machine learning, it cannot be specified how the model should use the variable (e.g., specify mediation or moderation in a path analysis). This point is particularly important given that we have many predictors in the model. It is possible that some predictors that had a causal effect on the outcome did not receive much credit because mediators of their effects were also included in the model, which may have reduced their direct effects. For example, the importance of IQ for predicting math achievement appeared to be relatively low, in contrast to some earlier findings (e.g., Sternberg et al., 2001). We hypothesize that in our analysis, the effect of IQ is

dampened due to the inclusion of prior achievement as a predictor, which is more proximal to the outcome variable than IQ. In our data, IQ and prior achievement were correlated (Pearson $r = .63$). Combined with the relatively small predictive effect of IQ on the outcome variable, this pattern of findings suggests that the effect of IQ was mediated through prior achievement. This interpretation aligns with evidence that measures of fluid intelligence are more predictive of academic success when children are younger and have less prior knowledge (Alloway & Alloway, 2010). Furthermore, the effect of motivation on math attainment may have been mediated by the choice of cognitive strategy, which may have reduced the main effect of motivation. Our predictive models are also unidirectional and thus do not explicitly model any of the reciprocal effects of math achievement that have been observed in prior studies (Arens et al., 2017; Hong et al., 2010; Marsh et al., 2022; Pekrun et al., 2017, 2023).

Although our methodology has many advantages, it also has its limitations. First, while our data aimed to incorporate most of the relevant variables related to math achievement based on expert inputs, certain predictors were not included. For example, some studies have indicated the importance of students' belief about effort in predicting math achievement (e.g., mindset; Blackwell et al., 2007), but we did not have these variables in the PALMA data. Therefore, it is possible that some results could change by including overlooked variables.

Second, some of the findings should be interpreted in the context of the German school system, which has a specific way of putting students in tracks. While we used an out-of-sample cross-validation procedure to test generalizability, this method ensures generalizability only for the population from which the sample was drawn. In other words, we cannot be sure that our findings generalize to other study populations (e.g., students from different countries with different cultural or institutional contexts). Future studies could examine whether and how the present approach can be helpful in cross-cultural research on students' achievement.

Finally, it is also important to recognize that, as is often the case with big data analyses in psychology and education, our analysis is a secondary data analysis. This means that some of the variable transformations, including the Rasch modeling of the outcome variable, had already been performed before we received the data. This has implications for potential information leakage, as strictly speaking, any data transformations should be made within the model training phase instead of being performed on the full data set. This is to prevent information about the test data from leaking into the training data. However, in the current investigation, any possible effects of leakage are expected to be minimal. The reason is that competence scores were estimated solely based on the student's answers to the mathematics test items, independent from the predictor variables. Therefore, information about the outcome is not confounded with information about the predictors. However, it is important to highlight this limitation in light of the ongoing discourse around transparency, "human-in-the-loop" overfitting (Hofman et al., 2017), and replicability of machine learning analyses (Gibney, 2022; Verstylen & Kording, 2023).

This analysis aimed to model all features simultaneously to see how important the different feature groups are when controlling for all other variables in the model. This "all-inclusive" approach was chosen as it most closely mimics the real-world scenario where many variables are interrelated and jointly contribute to achievement. We acknowledge that our approach is not the simplest, nor necessarily the most intuitive, but we believe it best captures the nuances that exist when predicting students' math achievement.

Conclusions

In summary, we utilized a longitudinal machine learning methodology to predict standardized math achievement scores in German secondary school students. Our approach enabled the simultaneous modeling of 105 different predictors and achieved high accuracy, validated on data unseen by the model. We also identified key temporal trends in the predictors' importance, most noticeably the increasing predictive power of prior achievement on future achievement. In addition, we found that data collected from surveys and cognitive tests added additional prediction accuracy beyond prior achievement and that their relative importance increased when predicting further into the future. Moreover, we found that classroom context was highly predictive when reported by the teachers (but not by the students). These results are particularly useful and applicable for researchers and practitioners who want to gauge the relative importance of many predictors of achievement, to identify when they matter most, and from whom to collect relevant data.

References

- Civitas. (2023). <https://www.civitaslearning.com/>
- Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology, 106*(1), 20–29. <https://doi.org/10.1016/j.jecp.2009.11.003>
- Altmann, A., Tološi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics (Oxford, England), 26*(10), 1340–1347. <https://doi.org/10.1093/bioinformatics/btq134>
- Arens, A. K., Frenzel, A. C., & Goetz, T. (2022). Self-concept and self-efficacy in math: Longitudinal interrelations and reciprocal linkages with achievement. *The Journal of Experimental Education, 90*(3), 615–633. <https://doi.org/10.1080/00220973.2020.1786347>
- Arens, A. K., Marsh, H. W., Pekrun, R., Lichtenfeld, S., Murayama, K., & Vom Hofe, R. (2017). Math self-concept, grades, and achievement test scores: Long-term reciprocal effects across five waves and three achievement tracks. *Journal of Educational Psychology, 109*(5), 621–634. <https://doi.org/10.1037/edu0000163>
- Au, Q., Herbinger, J., Stachl, C., Bischl, B., & Casalicchio, G. (2022). Grouped feature importance and combined features effect plot. *Data Mining and Knowledge Discovery, 36*(4), 1401–1450. <https://doi.org/10.1007/s10618-022-00840-5>
- Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*(3), 411–421. <https://doi.org/10.1097/01.psy.0000127692.23278.a9>
- Bailey, D. H., Watts, T. W., Littlefield, A. K., & Geary, D. C. (2014). State and trait effects on individual differences in children's mathematical development. *Psychological Science, 25*(11), 2017–2026. <https://doi.org/10.1177/0956797614547539>
- Baker, R. S., Esbenshade, L., Vitale, J., & Karumbaiah, S. (2023). Using demographic data as predictor variables: A questionable choice. *Journal of Educational Data Mining, 15*(2), 22–52. <https://doi.org/10.5281/zenodo.7702628>
- Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17. <https://doi.org/10.5281/zenodo.3554657>
- Bilal, M., Omar, M., Anwar, W., Bokhari, R. H., & Choi, G. S. (2022). The role of demographic and academic features in a student performance prediction. *Scientific Reports, 12*(1), Article 12508. <https://doi.org/10.1038/s41598-022-15880-6>
- Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition:

- A longitudinal study and an intervention. *Child Development*, 78(1), 246–263. <https://doi.org/10.1111/cdev.2007.78.issue-1>
- Blossfeld, H.-P., Schneider, T., & Doll, J. (2009). Methodological advantages of panel studies. Designing the new national educational panel study (NEPS) in Germany. *Journal for Educational Research Online*, 1(1), 10–32. <https://doi.org/10.25656/01:4554>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44(1), 108–132. <https://doi.org/10.1006/jmps.1999.1279>
- Cetintas, S., Si, L., Xin, Y. P., & Hord, C. (2009, July 1–3). *Predicting correctness of problem solving from low-level log data in intelligent tutoring systems*. 2nd International Conference on Educational Data Mining, Cordoba, Spain. <https://files.eric.ed.gov/fulltext/ED539041.pdf>
- Chakrapani, P., & Chitradevi, D. (2022, April 22–23). Academic performance prediction using machine learning: A comprehensive & systematic review. In *2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC), Chennai, India* (pp. 335–340). IEEE. <https://ieeexplore.ieee.org/document/9783512>
- Chen, T., & Guestrin, C. (2016, August 13–17). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, United States. (pp. 785–794). Association for Computing Machinery. <https://dl.acm.org/doi/pdf/10.1145/2939672.2939785>
- Christie, T., Jarratt, D. C., Olson, L. A., & Taijala, T. T. (2019). *Machine-learned school dropout early warning at scale*. <https://www.infnitecampus.com/pdf/Machinelearned-School-Dropout-Early-Warning-at-Scale.pdf>
- Cohausz, L., Tschalzev, A., Bartelt, C., & Stuckenschmidt, H. (2023). *Investigating the importance of demographic features for EDM-predictions*. International Educational Data Mining Society.
- Cox, C. R., Moscardini, E. H., Cohen, A. S., & Tucker, R. P. (2020). Machine learning for suicidology: A practical review of exploratory and hypothesis-driven approaches. *Clinical Psychology Review*, 82, Article 101940. <https://doi.org/10.1016/j.cpr.2020.101940>
- Dawson, S., Mirriahi, N., & Gasevic, D. (2015). Importance of theory in learning analytics in formal and workplace settings. *Journal of Learning Analytics*, 2(2), 1–4. <https://doi.org/10.18608/jla.2015.22.1>
- Daza, A., Guerra, C., Cervera, N., & Burgos, E. (2022). Predicting academic performance through data mining: A systematic literature. *TEM Journal*, 11(2), 939–949. <https://doi.org/10.18421/TEM>
- Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35(1), 13–21. <https://doi.org/10.1016/j.intell.2006.02.001>
- Deho, O. B., Joksimovic, S., Li, J., Zhan, C., Liu, J., & Liu, L. (2027). Should learning analytics models include sensitive attributes? Explaining the why. *IEEE Transactions on Learning Technologies*, 35(1), 13–21. <https://doi.org/10.1016/j.intell.2006.02.001>
- Deiningner, H., Lavelle-Hill, R., Parrisius, C., Pieronczyk, I., Colling, L., Meurers, D., Trautwein, U., Nagengast, B., & Kasneci, G. (2023). Can you solve this on the first try?—Understanding exercise field performance in an intelligent tutoring system. In *International Conference on Artificial Intelligence in Education* (pp. 565–576). Springer. https://link.springer.com/chapter/10.1007/978-3-031-36272-9_46
- Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Intergenerational class mobility in three Western European societies: England, France and Sweden. *The British Journal of Sociology*, 30(4), 415–441. <https://doi.org/10.2307/589632>
- Esbenshade, L., Baker, R. S., & Vitale, J. (2023). *From a prediction model to meaningful reports in school*. https://www.researchgate.net/publication/371911109_From_a_Prediction_Model_to_Meaningful_Reports_in_School
- Fan, Y., Saint, J., Singh, S., Jovanovic, J., & Gašević, D. (2021, April). A learning analytic approach to unveiling self-regulatory processes in learning tactics. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 184–195). Association for Computing Machinery. <https://dl.acm.org/doi/abs/10.1145/3448139.3448211>
- Frenzel, A. C., Goetz, T., Lüdtke, O., Pekrun, R., & Sutton, R. E. (2009). Emotional transmission in the classroom: Exploring the relationship between teacher and student enjoyment. *Journal of Educational Psychology*, 101(3), 705–716. <https://doi.org/10.1037/a0014695>
- Frenzel, A. C., Goetz, T., Pekrun, R., & Watt, H. M. (2010). Development of mathematics interest in adolescence: Influences of gender, family, and school context. *Journal of Research on Adolescence*, 20(2), 507–537. <https://doi.org/10.1111/jora.2010.20.issue-2>
- Gamazo, A., & Martínez-Abad, F. (2020). An exploration of factors linked to academic performance in Pisa 2018 through data mining techniques. *Frontiers in Psychology*, 11, Article 575167. <https://doi.org/10.3389/fpsyg.2020.575167>
- Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, 28, 68–84. <https://doi.org/10.1016/j.iheduc.2015.10.002>
- Geary, D. C., Nicholas, A., Li, Y., & Sun, J. (2017). Developmental change in the influence of domain-general abilities and domain-specific knowledge on mathematics achievement: An eight-year longitudinal study. *Journal of Educational Psychology*, 109(5), 680–693. <https://doi.org/10.1037/edu0000159>
- Gibney, E. (2022). *Could machine learning fuel a reproducibility crisis in science?* <https://doi.org/10.1038/d41586-022-02035-w>
- Goldberg, P., Sümer, Ö., Stürmer, K., Wagner, W., Göllner, R., Gerjets, P., Kasneci, E., & Trautwein, U. (2021). Attentive or not? Toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educational Psychology Review*, 33, 27–49. <https://doi.org/10.1007/s10648-019-09514-z>
- Gomes, J., Yassine, M., Worsley, M., & Blikstein, P. (2013). Analysing engineering expertise of high school students using eye tracking and multi-modal learning analytics. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Educational data mining 2013*. (pp. 375–377). International Educational Data Mining Society. <https://researchr.org/publication/edm-2013>
- Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science*, 17(1), 205–215. <https://doi.org/10.1177/1745691620984483>
- Hastie, T. (2020). Ridge regularization: An essential concept in data science. *Technometrics*, 62(4), 426–433. <https://doi.org/10.1080/00401706.2020.1791959>
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.
- Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- Hellas, A., Ithantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: A systematic literature review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education* (pp. 175–199). Association for Computing Machinery. <https://doi.org/10.1145/3293881.3295783>
- Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, revision: KFT 4-12+ R. Beltz-Test*.
- Henninger, M., Debelak, R., Rothacher, Y., & Strobl, C. (2023). Interpretable machine learning for psychological research: Opportunities and pitfalls. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000560>
- Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, 355(6324), 486–488. <https://doi.org/10.1126/science.aal3856>

- Hong, S., & Ho, H.-Z. (2005). Direct and indirect longitudinal effects of parental involvement on student achievement: Second-order latent growth modeling across ethnic groups. *Journal of Educational Psychology, 97*(1), 32–42. <https://doi.org/10.1037/0022-0663.97.1.32>
- Hong, S., Yoo, S.-K., You, S., & Wu, C.-C. (2010). The reciprocal relationship between parental involvement and mathematics achievement: Autoregressive cross-lagged modeling. *The Journal of Experimental Education, 78*(4), 419–439. <https://doi.org/10.1080/00220970903292926>
- Howard, T. C. (2019). *Why race and culture matter in schools: Closing the achievement gap in America's classrooms*. Teachers College Press.
- Issah, I., Appiah, O., Appiahene, P., & Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. *Decision Analytics Journal, 7*, Article 100204. <https://doi.org/10.1016/j.dajour.2023.100204>
- Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science, 15*(3), 809–816. <https://doi.org/10.1177/1745691620902467>
- Jaekel, A.-K., Göllner, R., & Trautwein, U. (2021). How students' perceptions of teaching quality in one subject are impacted by the grades they receive in another subject: Dimensional comparisons in student evaluations of teaching quality. *Journal of Educational Psychology, 113*(4), 770–783. <https://doi.org/10.1037/edu0000488>
- Khanna, L., Singh, S. N., & Alam, M. (2016, August 12–14). Educational data mining and its role in determining factors affecting students academic performance: A systematic review. In *2016 1st India International Conference on Information Processing (IICIP), Delhi, India* (pp. 1–7). IEEE. <https://ieeexplore.ieee.org/document/7975354>
- Kiray, S. A., Gok, B., & Bozkir, A. S. (2015). Identifying the factors affecting science and mathematics achievement using data mining methods. *Journal of Education in Science Environment and Health, 1*(1), 28–48. <https://doi.org/10.21891/jeseh.41216>
- Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology, 105*(3), 805–820. <https://doi.org/10.1037/a0032583>
- Lavelle-Hill, R., Goulding, J., Smith, G., Clarke, D. D., & Bibby, P. A. (2020). Psychological and demographic predictors of plastic bag consumption in transaction data. *Journal of Environmental Psychology, 72*, Article 101473. <https://doi.org/10.1016/j.jenvp.2020.101473>
- Lavelle-Hill, R., Smith, G., & Murayama, K. (2023). *Machine learning methods for large survey data in the social sciences: Challenges, solutions, and future directions*. OSF preprint. <https://doi.org/10.31219/osf.io/6xt82>
- Lee, K., & Bull, R. (2016). Developmental changes in working memory, updating, and math achievement. *Journal of Educational Psychology, 108*(6), 869–882. <https://doi.org/10.1037/edu0000090>
- Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences, 9*(15), Article 3093. <https://doi.org/10.3390/app9153093>
- Lin, X., & Powell, S. R. (2022). The roles of initial mathematics, reading, and cognitive skills in subsequent mathematics performance: A meta-analytic structural equation modeling approach. *Review of Educational Research, 92*(2), 288–325. <https://doi.org/10.3102/00346543211054576>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 4765–4774). Curran Associates. <https://papers.nips.cc/paper/7062-aunified-approach-to-interpreting-model-predictions.pdf>
- Marsh, H. W., Pekrun, R., & Lüdtke, O. (2022). Directional ordering of self-concept, school grades, and standardized tests over 5 years: New tripartite models juxtaposing within-and between-person perspectives. *Educational Psychology Review, 34*(4), 2697–2744. <https://doi.org/10.1007/s10648-022-09662-9>
- Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology, 111*(2), 331–353. <https://doi.org/10.1037/edu0000281>
- Martinez Abad, F., & Chaparro Caso López, A. A. (2017). Data-mining techniques in detecting factors linked to academic achievement. *School Effectiveness and School Improvement, 28*(1), 39–55. <https://doi.org/10.1080/09243453.2016.1235591>
- Matcha, W., Gašević, D., & Pardo, A. (2019). A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies, 13*(2), 226–245. <https://doi.org/10.1109/TLT.4620076>
- McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114*(2), 376–390. <https://doi.org/10.1037/0033-2909.114.2.376>
- McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research, 50*(5), 471–484. <https://doi.org/10.1080/00273171.2015.1036965>
- Molenaar, I., & Knoop-van Campen, C. (2017). Teacher dashboards in practice: Usage and impact. In É. Lavoué, H. Drachler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Eds.), *Data Driven Approaches in Digital Education. EC-TEL 2017. Lecture notes in computer science* (Vol. 10474, pp. 125–138). Springer. https://doi.org/10.1007/978-3-319-66610-5_10
- Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., & Bischl, B. (2020, July 18). General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Vienna, Austria* (pp. 39–68). Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-04083-2_4
- Muis, K. R., Sinatra, G. M., Pekrun, R., Winne, P. H., Trevors, G., Losenno, K. M., & Munzar, B. (2018). Main and moderator effects of refutation on task value, epistemic emotions, and learning strategies during conceptual change. *Contemporary Educational Psychology, 55*, 155–165. <https://doi.org/10.1016/j.cedpsych.2018.10.001>
- Murayama, K. (2022). A reward-learning framework of knowledge acquisition: An integrated account of curiosity, interest, and intrinsic-extrinsic rewards. *Psychological Review, 129*(1), 175–198. <https://doi.org/10.1037/rev0000349>
- Murayama, K., Pekrun, R., Lichtenfeld, S., & Vom Hofe, R. (2013). Predicting long-term growth in students' mathematics achievement: The unique contributions of motivation and cognitive strategies. *Child Development, 84*(4), 1475–1490. <https://doi.org/10.1111/cdev.2013.84.issue-4>
- Murayama, K., Pekrun, R., Suzuki, M., Marsh, H. W., & Lichtenfeld, S. (2016). Don't aim too high for your kids: Parental overaspiration undermines students' learning in mathematics. *Journal of Personality and Social Psychology, 111*(5), 766–779. <https://doi.org/10.1037/pspp0000079>
- Nadaf, A., Eliëns, S., & Miao, X. (2021). Interpretable-machine-learning evidence for importance and optimum of learning time. *International Journal of Information and Education Technology (IJJET), 11*(10), 444–449. <https://doi.org/10.18178/ijjet.2021.11.10.1548>
- Nadaf, A., Monroe, S., Chandran, S., & Miao, X. (2022, July). Learning factors for TIMSS math performance evidenced through machine learning in the UAE. In *International Conference on Artificial Intelligence in Education Technology* (pp. 47–66). Springer Nature Singapore.
- Noetel, M., Parker, P., Dicke, T., Beauchamp, M. R., Ntoumanis, N., Hulteen, R. M., Diezmann, C., Yeung, A., Ahmadi, A., Vasconcellos, D., & Mahoney, J. (2023). Prediction versus explanation in educational psychology: A cross-theoretical approach to using teacher behaviour to predict student engagement in physical education. *Educational Psychology Review, 35*(3), Article 73. <https://doi.org/10.1007/s10648-023-09786-6>
- Orth, U., Meier, L. L., Bühler, J. L., Dapp, L. C., Krauss, S., Messerli, D., & Robins, R. W. (2024). Effect size guidelines for cross-lagged effects.

- Psychological Methods*, 29(2), 421–433. <https://doi.org/10.1037/met0000499>
- Paquette, L., Ocumpaugh, J., Li, Z., Andres, A., & Baker, R. (2020). Who's learning? Using demographics in EDM research. *Journal of Educational Data Mining*, 12(3), 1–30. <https://doi.org/10.5281/zenodo.4143612>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011a). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011b). *Scikit-learn: Machine learning in Python. Iterative imputer*. <https://scikit-learn.org/stable/modules/impute.html#iterative-imputer>
- Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., & Goetz, T. (2017). Achievement emotions and academic performance: Longitudinal models of reciprocal effects. *Child Development*, 88(5), 1653–1670. <https://doi.org/10.1111/cdev.2017.88.issue-5>
- Pekrun, R., Marsh, H. W., Suessbach, F., Frenzel, A. C., & Goetz, T. (2023). School grades and students' emotions: Longitudinal models of within-person reciprocal effects. *Learning and Instruction*, 83, Article 101626. <https://doi.org/10.1016/j.learninstruc.2022.101626>
- Pekrun, R., Murayama, K., Marsh, H. W., Goetz, T., & Frenzel, A. C. (2019). Happy fish in little ponds: Testing a reference group model of achievement and emotion. *Journal of Personality and Social Psychology*, 117(1), 166–185. <https://doi.org/10.1037/pspp0000230>
- Pekrun, R., Vom Hofe, R., Blum, W., Frenzel, A. C., Goetz, T., & Wartha, S. (2007). *Development of mathematical competencies in adolescence: The PALMA longitudinal study*. Konstanzer Online-Publikations-System. <https://kops.uni-konstanz.de/server/api/core/bitstreams/12dd3457-45e4-4f91-a43f-493515a52e95/content>
- Picciano, A. G. (2012). The evolution of big data and learning analytics in American higher education. *Journal of Asynchronous Learning Networks*, 16(3), 9–20. <https://doi.org/10.24059/ojs.v16i3.267>
- PISA. (2018). *PISA technical report: Chapter 16 scaling procedures and construct validation of context questionnaire data*. https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018_Technical-Report-Chapter-16-Background-Questionnaires.pdf
- Psyridou, M., Koponen, T., Tolvanen, A., Aunola, K., Lerkkanen, M.-K., Poikkeus, A.-M., & Torppa, M. (2024). Early prediction of math difficulties with the use of a neural networks model. *Journal of Educational Psychology*, 116(2), 212–232. <https://doi.org/10.1037/edu0000835>
- Rawson, K., Stahovich, T. F., & Mayer, R. E. (2017). Homework and achievement: Using smartpen technology to find the connection. *Journal of Educational Psychology*, 109(2), 208–219. <https://doi.org/10.1037/edu0000130>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). ACM. <https://dl.acm.org/doi/pdf/10.1145/2939672.2939778?>
- Rogers, T., Dawson, S., & Gasevic, D. (2016). Learning analytics and the imperative for theory driven research. In *The SAGE handbook of e-learning research* (2nd ed., pp. 232–250). SAGE Publications. <https://doi.org/10.4135/9781473955011.n12>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., & Datta, D. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, 117(15), 8398–8403. <https://doi.org/10.1073/pnas.1915006117>
- Scheinost, D., Noble, S., Horien, C., Greene, A. S., Lake, E. M., Salehi, M., Gao, S., Shen, X., O'Connor, D., Barron, D. S., & Yip, S. W. (2019). Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage*, 193, 35–45. <https://doi.org/10.1016/j.neuroimage.2019.02.057>
- Sekeroglu, B., Abiyev, R., Ilhan, A., Arslan, M., & Idoko, J. B. (2021). Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies. *Applied Sciences*, 11(22), Article 10907. <https://doi.org/10.3390/app112210907>
- Selzam, S., Krapohl, E., Von Stumm, S., O'Reilly, P. F., Rimfeld, K., Kovas, Y., Dale, P., Lee, J., & Plomin, R. (2017). Predicting educational achievement from DNA. *Molecular Psychiatry*, 22(2), 267–272. <https://doi.org/10.1038/mp.2016.107>
- Shapley, L. (1997). 7. A value for n-person games. Contributions to the theory of games II (1953) 307-317. In H. Kuhn (Ed.), *Classics in game theory* (pp. 69–79). Princeton University Press. <https://doi.org/10.1515/9781400829156-012>
- Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, 46(5), 30–40. <https://eric.ed.gov/?id=EJ950794>
- Steinmayr, R., & Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learning and Individual Differences*, 19(1), 80–90. <https://doi.org/10.1016/j.lindif.2008.05.004>
- Stekhoven, D. J., & Bühlmann, P. (2012). Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118. <https://doi.org/10.1093/bioinformatics/btr597>
- Sternberg, R. J., Grigorenko, E. L., & Bundy, D. A. (2001). The predictive value of IQ. *Merrill-Palmer Quarterly*, 47(1), 1–41. <https://doi.org/10.1353/mpq.2001.0005>
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4), 323–348. <https://doi.org/10.1037/a0016973>
- Talsma, K., Schüz, B., Schwarzer, R., & Norris, K. (2018). I believe, therefore I achieve (and vice versa): A meta-analytic cross-lagged panel analysis of self-efficacy and academic performance. *Learning and Individual Differences*, 61, 136–150. <https://doi.org/10.1016/j.lindif.2017.11.015>
- Tamura, A., Ishii, R., Yagi, A., Fukuzumi, N., Hatano, A., Sakaki, M., Tanaka, A., & Murayama, K. (2022). Exploring the within-person contemporaneous network of motivational engagement. *Learning and Instruction*, 81, Article 101649. <https://doi.org/10.1016/j.learninstruc.2022.101649>
- Usami, S., Murayama, K., & Hamaker, E. L. (2019). A unified framework of longitudinal models to examine reciprocal relations. *Psychological Methods*, 24(5), 637–657. <https://doi.org/10.1037/met0000210>
- Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67. <https://doi.org/10.18637/jss.v045.i03>
- Van Lissa, C. J. (2022). Developmental data science: How machine learning can advance theory formation in developmental psychology. *Infant and Child Development*, 32(6), Article e2370. <https://doi.org/10.1002/icd.2370>
- Verstynen, T., & Kording, K. P. (2023). Overfitting to ‘predict’ suicidal ideation. *Nature Human Behaviour*, 7(5), 680–681. <https://doi.org/10.1038/s41562-023-01560-6>
- von Stumm, S., Smith-Woolley, E., Ayorech, Z., McMillan, A., Rimfeld, K., Dale, P. S., & Plomin, R. (2020). Predicting educational achievement from genomic measures and socioeconomic status. *Developmental Science*, 23(3), Article e12925. <https://doi.org/10.1111/desc.v23.3>
- Wong, B. T.-m., & Li, K. C. (2020). A review of learning analytics intervention in higher education (2011–2018). *Journal of Computers in Education*, 7(1), 7–28. <https://doi.org/10.1007/s40692-019-00143-7>
- Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining

- and theory. *Computers in Human Behavior*, 47, 168–181. <https://doi.org/10.1016/j.chb.2014.09.034>
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–1122. <https://doi.org/10.1177/1745691617693393>
- Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, 81(2), 267–301. <https://doi.org/10.3102/0034654311405999>
- Yoo, J. E. (2018). TIMSS 2011 student and teacher predictors for mathematics achievement explored and identified via elastic net. *Frontiers in Psychology*, 9, Article 317. <https://doi.org/10.3389/fpsyg.2018.00317>
- Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should college dropout prediction models include protected attributes? In *Proceedings of the Eighth ACM Conference on Learning@ Scale* (pp. 91–100). Association for Computing Machinery.
- Zhang, W., Zhou, Y., & Yi, B. (2019). An interpretable online learner's performance prediction model based on learning analytics. In *Proceedings of the 11th International Conference on Education Technology and Computers* (pp. 148–154). Association for Computing Machinery.
- Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Elsevier.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>

Received August 15, 2023

Revision received January 29, 2024

Accepted January 30, 2024 ■