Who Did What to Succeed? Individual Differences in Which Learning Behaviors Are Linked to Achievement

Hannah Deininger
Hector Research Institute of
Education Sciences and Psychology
University of Tübingen
Tübingen, Germany
hannah.deininger@uni-tuebingen.de

Cora Parrisius
Karlsruhe University of Education
Karlsruhe, Germany
cora.parrisius@ph-karlsruhe.de

Rosa Lavelle-Hill
Department of Psychology &
Copenhagen Center for Social Data
Science (SODAS)
University of Copenhagen
Copenhagen, Denmark
rla@psy.ku.dk

Detmar Meurers
Language and AI in Education Lab
Leibniz Institut für Wissensmedien
(IWM)
Tübingen, Germany
d.meurers@iwm-tuebingen.de

Ulrich Trautwein
Hector Research Institute of
Education Sciences and Psychology
University of Tübingen
Tübingen, Germany
ulrich.trautwein@uni-tuebingen.de

Benjamin Nagengast
Hector Research Institute of
Education Sciences and Psychology
University of Tübingen
Tübingen, Germany
benjamin.nagengast@uni-tuebingen.de

Gjergji Kasneci
Responsible Data Science
Technical University of Munich
Munich, Germany
gjergji.kasneci@tum.de

Abstract

It is commonly assumed that digital learning environments such as intelligent tutoring systems facilitate learning and positively impact achievement. This study explores how different groups of students exhibit distinct relationships between learning behaviors and academic achievement in an intelligent tutoring system for English as a foreign language. We examined whether these differences are linked to students' prior knowledge, personality traits, and motivation. We collected behavioral trace data from 507 German seventh-grade students during the 2021/22 school year and applied machine learning models to predict English performance based on learning behaviors (best-performing model's R^2 = .41). To understand the impact of specific behaviors, we applied the explainable AI method SHAP and identified three student clusters with distinct learning behavior patterns. Subsequent analyses revealed that these clusters also varied in prior knowledge and motivation: one with high prior knowledge and average motivation, another with low prior knowledge and average motivation, and a third with both low prior knowledge and low motivation. Our findings suggest that learning behaviors are linked differently to academic success across students and are closely tied to their prior knowledge and motivation. This hints towards the importance of personalizing learning systems to support individual learning needs better.

LAK 2025, Dublin, Ireland
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706571

CCS Concepts

• **Applied computing** → **Interactive learning environments**; *E-learning*; *Computer-managed instruction*; • **Computing methodologies** → *Supervised learning by regression.*

Keywords

Learning Analytics, Behavioral Trace Data, Academic Performance, Interindividual Differences

ACM Reference Format:
Hannah Deininger, Cora Parrisius, Rosa Lavelle-Hill, Detmar Meurers, Ulrich Trautwein, Benjamin Nagengast, and Gjergji Kasneci. 2025. Who Did What to Succeed? Individual Differences in Which Learning Behaviors Are Linked to Achievement. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3706468.3706571

1 Introduction

Students worldwide increasingly engage with digital learning environments [48], including intelligent tutoring systems (ITS), which are assumed to facilitate learning and improve achievement [30]. These systems allow researchers to collect fine-grained interaction data, known as behavioral trace data, that capture the learning process in real time. Analyzing these data enables studying learning in ecologically valid settings and predicting student achievement based on learning behaviors [1].

Previous research has already established a link between learning behavior and learning outcomes [11]. At the same time, educational psychology highlights the role of person-level factors, such as motivation or personality, in influencing learning behavior and achievement [20, 50]. These relationships have been empirically validated using self-report data, enabling researchers to define direct, mediating, and moderating effects between person-level factors, learning behavior, and achievement [23, 62]. With the introduction of behavioral trace data, the focus shifted from traditional statistical models to more complex machine learning (ML) models [1]. Although ML models excel at handling large amounts of highly intercorrelated data [12], they prioritize the predictive power of entire models instead of the predictive value and selection of individual features.

Even though prior research has linked learning behaviors to achievement [7, 16], inconsistencies remain about whether specific learning behaviors, such as time spent on learning activities, are positively or negatively linked to achievement [26]. This variability might stem from individual differences in person-level factors, such as prior knowledge, personality, and motivation, which likely moderate the effectiveness of learning behaviors [15, 34, 50]. For example, the amount of practice might be more critical for low-prior-knowledge students than for high-prior-knowledge ones. Similarly, re-practicing a task might be more predictive for low-conscientious students, as this behavior might be less typical for them but could effectively support the learning outcome. Thus, there is reason to assume that the relevance of the learning behaviors (e.g., the number of tasks submitted) for later performance differs among students depending on their prior knowledge, motivation, and personality characteristics. Integrating strong ML prediction models with insights regarding the influence of person-level factors on the link between learning behavior and achievement could deepen our understanding of this relationship. Our study presents a novel approach incorporating explainable AI (XAI) and clustering to re-introduce the theory-aligned and deliberate investigation of learning behaviors in predicting learning outcomes in an ML paradigm. More precisely, we will examine potential group differences in the relationship between learning behaviors and achievement.

2 Theoretical Background

2.1 Capturing Learning Behavior with Behavioral Trace Data

Digital learning environments, such as learning management systems in higher education or ITS in schools, have become widespread [48]. These systems collect data on how students interact with the system, providing insights into learning behaviors that are more granular and less limited by subjective bias than traditional self-reports [7]. Behavioral trace data typically consist of timestamped events, capturing students' progress through the system. These raw event logs are usually further processed into aggregated features, so-called learning behavior indicators [4]. Determining which indicators best capture learning remains debated, but efforts have been made to combine insights from learning analytics and educational psychology to create accurate and theory-based indicators [17, 22]. To adequately capture learning behavior, theory must inform the data preprocessing, analysis, and interpretation processes [61].

2.2 Learning Behavior and Its Link to Achievement

Various studies have shown that students' learning behaviors, as captured by behavioral trace data, impact their later achievement [25] (see [7] for an example with learning management data and [16, 27] for an example with ITS data).

Besides predicting learning outcomes based on learning behavior indicators, researchers have also sought to understand learning efficacy, i.e., what kind of learning behaviors are especially useful for learning [51]. Previous work mainly relies on clustering approaches that group students in specific learning behavior types and link these to learning outcomes [32, 56]. For example, Schütt et al. [51] used clustering to extract five student groups based on the average time between clicks associated with different learning gains in their system. Among others, they identified a group with low learning gains, fewer actions, and a short time between them, which they labeled few-shot rushers.

2.3 The Interaction Between Person-Level Factors and Learning Behaviors

While the fact that learning behaviors are generally relevant for learning outcomes is undisputed, debates remain about the direction and strength of this relationship. Some behaviors, such as time spent on tasks, have shown varying effects on learning outcomes across studies [26]. Educational psychology offers a potential explanation: Person-level factors, such as prior knowledge, motivation, and personality, may moderate the relationship between learning behaviors and achievement [50]. That is, depending on the person, the same action might mean different things and, in consequence, might or might not be relevant to their learning process. For example, students with high prior knowledge may benefit only a little from practicing many tasks, whereas practicing a lot is much more helpful for students with low prior knowledge. This may also be true for students' motivation or personality: Spending more time on learning tasks might be positively associated with learning outcomes for highly motivated students (indicating effortful work) but negative for students with low motivation (indicating unfocused work) [23]. The expectancy-value framework [19] posits that students' expectancies and values as indicators of their motivation drive their learning behaviors and achievement. This calls for investigating the relationship between trace-data-based learning behaviors and achievement on the overall sample level while also considering learners' preconditions and how these might affect this relationship.

Both educational psychology and learning analytics stress the need to incorporate such interaction terms, i.e., to investigate who shows a specific behavior and understand its impact on achievement. In the learning analytics community, studies have already shown that learning behavior itself is associated with person-level factors, such as prior knowledge [55], motivation [15], or personality traits [34]. Additionally, studies showed that person-level factors impact learning success above and beyond learning behavior [11]. Dispositional learning analytics, for example, combines behavioral data with self-report data to better understand the factors influencing learning outcomes [57]. However, within these models, the interaction between person-level factors and learning behaviors has not been explicitly investigated so far. In educational psychology, Efklides [20] described this interaction in the context of self-regulated learning (SRL) theory. She stated that each learner brings their individual differences (e.g., regarding motivation) in the learning situation that affect how their learning behavior is linked to their later achievement, which was investigated in multiple studies (see [21] for an overview).

3 Research Questions

Modeling the interaction effect of person-level factors on the association between learning behaviors and achievement is not straightforward. One reason is that behavioral trace data build up large, highly inter-correlated feature sets traditional regression models cannot handle. ML models, on the other hand, may solve this issue but cannot be used to specify expected moderation effects explicitly. To still be able to investigate the interaction of person-level factors and learning behavior on achievement, we propose a novel approach combining ML with XAI and clustering. Previous research has used clustering approaches on behavioral data (e.g., [51]) and local explanations for similar objectives (e.g., [3, 28]). However, the major contribution of this paper lies in its novel combination of both approaches. By training an ML model on learning behavior indicators and generating SHAP explanations [33] for this relationship, we obtained a quantification of the relevance per learning behavior indicator for the prediction. While SHAP is designed for local explanations, our methodology mitigates this by clustering the SHAP values themselves, which results in clusters of students for whom the link between specific learning behavior indicators and the outcome is similar. Furthermore, we acknowledge that the learning behavior-achievement relationship may vary depending on the individual. Thus, we additionally investigated the potential interaction of person-level factors with the learning behavior-achievement link. This three-step process allowed us to identify subgroups with different individual preconditions, which is crucial for achieving personalization beyond a one-size-fits-all model.

We applied this novel multi-step approach to learning behavior data from an ITS for learning English as a foreign language in German seventh-grade classes. We asked the following research questions to (a) identify student clusters and (b) describe the clusters based on learning behaviors and person-level factors:

(1) Can we identify clusters of students for whom similar sets of learning behavior indicators are relevant for predicting later English proficiency?
(2) Do the students from the identified clusters differ regarding their displayed learning behaviors?
(3) Do the students from the identified clusters differ regarding their prior knowledge, motivation, or personality?

4 Method

4.1 Data Set

This study draws on data from a cluster-randomized controlled field trial, the Interact4School study [39], investigating foreign language learning in German secondary schools through individualized practice using the ITS FeedBook [14, 49]. For transparency, proficiency tests, questionnaire items, and hyper-parameter grids for all trained ML models are listed on the OSF: https://osf.io/2v9gf/?view_only=2154581235ac46bb8ce719e82168f793.

4.1.1 Sample. A total of 618 students from 24 classes across seven public academic-track schools in Germany used the ITS for English as a foreign language during the 2021/2022 school year. To focus specifically on students' learning behavior within the ITS, we included only those who actively used it, resulting in a final sample of $N = 507$ (55.8% female, $M_{age} = 12.50$ years, $SD_{age} = 0.41$). This participant reduction occurred because one teacher did not incorporate the ITS into their teaching during the study period. Additionally, up to five students per class did not use the ITS, although their classmates did.

Among the students, 35.1% had a migration background (i.e., either they or one of their parents were born outside Germany). 80.9% of the students had at least one parent with a university entrance qualification. This is approximately representative of German academic-track schools [18]. 4.2% of the students were native English speakers.

4.1.2 Procedure and Study Design. During the 2021/2022 school year, participating teachers and their students used the ITS FeedBook [49] to complement an English as a foreign language textbook for seventh graders in academic-track schools. The ITS consists of four thematic units, so-called task cycles, each containing individual tasks that focus on specific grammatical constructs, such as the simple past, and vocabulary [9].

Classes using the ITS were randomly assigned to one of three ITS versions, each with a different number of motivational elements. Additionally, students within each class were randomly assigned to one of two groups, differing in the granularity of feedback provided. By including study conditions and group information in our feature set, we controlled for all design factors that might affect our analyses.

Trained instructors collected self-report survey and English proficiency test data at the start of the school year (before the ITS was introduced; T1) and several weeks later, immediately after task cycle 1 (T2). Behavioral trace data was gathered via the ITS system. For this study, we focused on data from task cycle 1 to minimize data reduction, as there was ongoing class dropout throughout the school year.

4.1.3 Learning Behavior Indicators. The used FeedBook version contained 301 tasks (78 in task cycle 1). Each task consisted of five to ten items, typically including a blank that students were required to fill in—the task field. Other common task formats include multiple-choice and drag-and-drop. Students could attempt each task field multiple times, resulting in one or more attempts per task field. After each attempt, feedback was provided, indicating whether the response was correct or incorrect. Students in the enriched feedback group received additional, more detailed feedback. Throughout task cycle 1, students' interactions with FeedBook resulted in around 222,300 events. Each saved event consisted of the event message, the corresponding user, a timestamp, and the associated context (i.e., the task and task field related to the event). These raw events reflected students' practicing behavior in the system but required the engineering of indicators.

Our data extraction process was informed by existing literature on learning behavior and student interviews about their use of the ITS. A detailed description of the feature engineering process can be found in [17]. In this prior work, we focused on engineering theory-aligned behavioral indicators that describe the frequency, duration, density, and distribution of students' learning behavior in the ITS. The current study uses the resulting feature set as input for all trained ML models (see Table 1).[1] To account for class-level dependencies in the data, we introduced M and SD per class as features and group-mean centered all indicators at the class mean (in line with [16]). By one-hot encoding multi-categorical features, the dataset expanded to 221 features.

4.1.4 English Proficiency Score. Students' performance on an English language test at T2 (the achievement outcome to be predicted) was measured using a curricular-valid English proficiency test. This test focused on grammar topics included in task cycle 1: simple past, gerund, and modal verbs. It comprised 30 items where students needed to complete a sentence (e.g., "Yesterday I _____ (forget) my glasses at home.") Two items were removed due to ambiguous wording, resulting in 28 items (see [40] for more information).

We computed a sum score for each student (1 point per correctly completed sentence). The post-test scores ranged between 0 and 28 points with acceptable internal consistency (KR-20 = .85). We standardized the proficiency scores to ensure all variables were on a consistent scale.

4.1.5 Person-Level Factors. A general English proficiency test assessed students' prior English knowledge at T1 (i.e., a C-test; [44]) consisting of 125 sentence-completion-items. As for the post-test, we computed sum scores ranging from 0 to 125 with good internal consistency (KR-20 = .93) and also standardized the prior knowledge scores.

Students' motivation and personality were assessed at T1 using multiple self-report items, each rated on a 4-point response scale ranging from 1 (completely disagree) to 4 (completely agree). To measure subject-related expectancies, we used four items focused on students' English competence beliefs [24] and one item assessing satisfaction with perceived competence (e.g., "I am satisfied with my performance in English," adapted from the Intrinsic Motivation Inventory; [35]). Students' English value beliefs were measured using the value beliefs scale developed by Gaspard et al. [24]. It measures intrinsic value (three items, e.g., "English is fun to me"), attainment value (four items, e.g., "To be good at English means a lot to me"), utility value (four items, e.g., "English skills can be used well in everyday life and leisure time"), and cost (seven items, e.g., "Doing English is exhausting") based on expectancy-value theory [19]. All scales showed acceptable to good internal consistency (subject-related expectancies: α = .87; intrinsic value: α = .90; attainment value: α = .78; utility value: α = .68; cost: α = .82).

To assess students' conscientiousness and openness, we utilized the corresponding scales from the short version of the Big Five Inventory (BFI-K; German translation by [45]), consisting of four items for conscientiousness (e.g., "I do my tasks thoroughly") and five items for openness (e.g., "I am interested in many things.") Both scales demonstrated acceptable internal consistency (conscientiousness: α = .71; openness: α = .67).

We calculated factor scores for all scales by fitting one confirmatory factor analysis (CFA) model for all scale variables. We used a single latent variable each to capture conscientiousness, openness, and the expectancy component of motivation. Additionally, we utilized a bi-factor model for the value component of motivation, which simultaneously models both specific factors (i.e., specific intrinsic value, specific attainment value, specific utility value, and specific cost value) and a general factor that captures the shared variance among all items across the four value components [41]. Consequently, the specific factors represent variance unique to the specific value component and not shared among all four components. The model fit was acceptable ($RMSEA$ = $.048, 90\%CI[.039 - .046], SRMR = .059$). This variable set built the basis for the analysis of RQ3. The significance level for all analyses was set at $\alpha < .05$.

4.1.6 Outlier Handling for Learning Behavior Indicators. In educational log data, distinguishing between plausible and implausible values requires domain-specific knowledge and is influenced by the objectives of the analysis [47]. Consequently, all outlier values (i.e., those below the fifth or above the 95-th quantiles; [2]) were carefully examined for plausibility, namely, if these values were deemed impossible due to system errors or implausible due to off-task behavior (e.g., item response times exceeding 2 minutes for relatively short items). Our analysis determined that all outlier values were implausible. To address this, we employed winsorization (i.e., replacing extreme values with less extreme ones; [60]) with a trim quantile of .10. This approach allowed us to retain implausible outliers that were still valid by replacing them with reasonable reference values, thereby ensuring accurate data modeling, minimizing the impact of extreme values on results, and preventing significant data loss compared to the exclusion of participants with outlier values. Additionally, this method preserves the interpretability of variables, avoiding distortion from factors such as off-task behavior. Table 1 provides details on which variables were winsorized.

4.1.7 Missing Data. We observed missing data in students' self-reports at both measurement time points due to either student absences or non-responses to specific items (ranging from 2-14%). There were no missing data for the C-test or the English proficiency post-test. Although we only included students who actively used FeedBook during task cycle 1, some students had missing values for specific behavioral indicators, such as when they did not exhibit certain behaviors. Since the implementations of the ML algorithms used, except for extreme gradient boosting (XGBoost), could not handle missing data, K-Nearest Neighbours (KNN) imputation was employed to fill in missing values for all other algorithms [42, 54].

4.2 Analysis Pipeline

The first step as a basis for all further analyses was the generation of a high-quality prediction model using the extracted learning behavior indicators to predict the post-test English proficiency score (see Figure 1). Since we were dealing with tabular data, our

[1] Both studies work with the same set of behavioral indicators and similar survey variables. However, the previous research focused on bridging the gap between educational theories and learning analytics. In contrast, the current study investigates group differences in learning behavior's link to achievement. Different analytical approaches were chosen in each case.

Table 1: Learning behavior indicators engineered from the system logs. Based on these variables, class-level aggregations (i.e., M, SD) and group-mean centered variables on the class mean were calculated.

Behavioral Indicators	Winsorized?
No. of days worked with the ITS	–
Average learning sessions duration	✓
Average time on task	–
Average time on feedback	✓
Average time between the appearance of feedback and the subsequent attempt	–
Average time per task field	✓
Average response time per attempt	✓
Average response time for first attempt	✓
Average response time for second attempt	✓
Average response time for last attempt	✓
Average time between tasks within learning sessions	–
Ratio of time spent on task to average sample time spent on task	–
If clicked on a question mark to request feedback	–
No. of clicks on grammar help page	–
Proportion of task fields correct at submission	–
Proportion of used opportunities to correct	–
Proportion of correct at first try	–
Proportion of uptakes (incorrect attempt ->feedback ->correct attempt)	–
Proportion of tasks only attempted but not submitted	–
If tasks were practiced again in practice mode	–
Proportion of more practice tasks	–
Proportion of capstone tasks	–
If tasks were viewed after submission	–
Proportion of task fields submitted correctly after multiple incorrect attempts	–
Average no. of attempts per task field	–
No. of submitted tasks in relation to the class mean	–
Proportion of tasks submitted completely	–
Average proportion of task fields submitted per task	–
No. of tasks attempted	–
Submission rate (average no. of submissions per day)	–
Average sequence length	✓
Average no. of task fields per learning session	–
If there was mainly straight progress through task fields	–
If tasks were mainly finished with a last brief attempt	–
If there were mainly correct first-try attempts at the end of a learning session	–
If tasks were completed right before school	–
Average finish time difference relative to first in class per task	–
Average interval until the median of the class has been submitted	–
Distribution of learning sessions over time	✓
Proportion of active practicing days among all days in task cycle 1	–
Proportion of active days in relation to the class mean of active days	–

focus was on ML approaches varying in complexity that are known to handle this type of data well [10], such as linear regression, lasso regression [58], support vector regression [5], decision tree [46], random forest [12], and XGBoost [13]. To train all models, we first split the data set into train and test set based on an 80:20 partition. All models were then trained on the train set with 10-fold cross-validation using grid search for hyper-parameter tuning [2]. To measure the models' performance, we compared the predictive power of all models to a mean baseline and assessed explained variance, RMSE, and MAE to identify the most predictive model (Table 2). All preprocessing steps, as well as the ML and XAI analysis, were conducted with Python 3.8 using `scikit-learn` 1.3.2 [42], `xgboost` 1.6.2 [13], and `shap` 0.41.0 [33]. Cluster comparisons were conducted with R version 4.3.3 [43].

The best-performing XGBoost model was then used to generate explanations for the test set predictions using the XAI method

[2] See OSF for hyper-parameter grids: https://osf.io/2v9gf/?view_only=2154581235ac46bb8ce719e82168f793

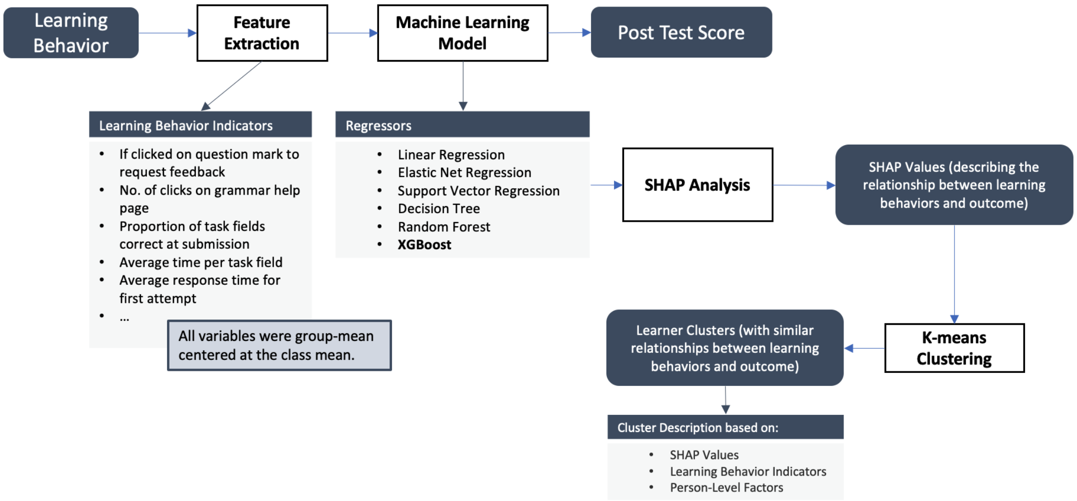
Learning Behavior
Feature Extraction
Machine Learning Model
Post Test Score
Learning Behavior Indicators
If clicked on question mark to request feedback
No. of clicks on grammar help page
Proportion of task fields correct at submission
Average time per task field
Average response time for first attempt
...
All variables were group-mean centered at the class mean.
Regressors
Linear Regression
Elastic Net Regression
Support Vector Regression
Decision Tree
Random Forest
XGBoost
SHAP Analysis
SHAP Values (describing the relationship between learning behaviors and outcome)
K-means Clustering
Learner Clusters (with similar relationships between learning behaviors and outcome)
Cluster Description based on:
SHAP Values
Learning Behavior Indicators
Person-Level Factors

Figure 1: Analysis pipeline.

SHAP [3] [33]. SHAP generates post hoc, local explanations in the form of SHAP values. For each student, the SHAP values quantify the impact of each learning behavior indicator on the outcome prediction. More formally, they describe the impact of each variable on the outcome's difference from the average prediction, the so-called base value [36]. A high absolute SHAP value indicates higher importance for the prediction outcome of that specific learning behavior. In addition, the valence of the SHAP value describes the direction of influence, with positive values indicating an increase and negative values a decrease in the model's prediction [36]. In other words, a positive SHAP value indicates that a specific learning behavior increases the predicted post-test score, a SHAP value close to zero indicates no change in the predicted post-test score, and a negative SHAP value indicates that the learning behavior reduces the predicted post-test score in relation to the base value.

Table 2: Performance metrics for all trained machine learning models on the hold-out test set. The lowest error is in italics.

	MAE	RMSE	Explained Variance
Mean Baseline	8.35	10.09	0.00
Linear Regression	4.38	5.37	0.25
Elastic Net Regression	4.22	5.21	0.29
Decision Tree	3.94	5.14	0.31
Support Vector Regression	4.38	5.47	0.24
Random Forest	*3.77*	4.79	0.39
XGBoost	3.83	*4.74*	*0.41*

We considered SHAP a suitable approach for explaining the prediction because it is model-agnostic, which means it can generate explanations for any ML model. This was important since we investigated the predictive power of multiple ML algorithms and wanted to ensure the explainability for all of them. Further, it generates local explanations per instance, enabling comparisons between individual students. Additionally, these local explanations can be aggregated to gain insights at a global level (e.g., sample or group level) [33], which helps to understand the learning mechanisms across groups of students, as we want to investigate in this study. Furthermore, SHAP can account for high correlations among input features by distributing their shared impact across the correlated features [4] [33], which is essential because of highly inter-correlated features in our dataset. To ensure the stability of the generated explanations, we generated SHAP values for 100 bootstrapped test set samples and assessed standard deviations across all instances and explanations. With an average standard deviation of < 0.001, we assume the explanations to be stable.

The generated SHAP values – describing the impact of each of the learning behaviors on the prediction – were then used to cluster the students using k-means as implemented in `scikit-learn` [42] (RQ1). That is, students with similar SHAP values per learning behavior indicator were clustered together, assuming that for students of the same cluster, their learning behaviors were similarly linked to their later predicted performance. In line with other studies investigating behavioral trace data [51, 59], we chose k-means clustering because it is straightforward to control and analyze. Further, it can deal with multiple continuous variables, has good computational performance, and produces easily interpretable results [52]. To obtain the optimal number of clusters, we used silhouette score analysis [53]. Again, to ensure the stability of the derived number of clusters, we repeated the silhouette score analysis based on 100 bootstrapped SHAP value samples. Since the standard deviation of the silhouette scores was low across all bootstrapped samples ($SD < 0.001$), we assumed the preferred number of clusters to be stable.

[3] Based on the true to the data approach.

[4] For the true-to-the-data approach.

As a next step, we compared the average post-test score (outcome) between clusters and described a typical student per cluster based on their displayed learning behaviors (RQ2). Finally, to better understand the student clusters (i.e., to examine potential interactions of person-level factors with learning behaviors on achievement), we compared them concerning their prior knowledge, motivation, and personality levels using a MANOVA (RQ3). These variables were not used as features within the ML models.

5 Results

5.1 RQ1: Can we Identify Clusters of Students for Whom Similar Sets of Learning Behavior Indicators are Relevant for Predicting Later English Proficiency?

Silhouette score analyses across 100 bootstrapped samples resulted in acceptable scores for a two- ($M = 0.41, SD < .001$) and three-cluster solution ($M = 0.34, SD < .001$). Since the two-cluster solution seemed to mainly focus on differences in prior knowledge without much practical relevance, we favored the three-cluster solution.

Overall, roughly the same indicators were relevant across clusters. Still, the valence and strength of the SHAP values differed among clusters. For example, the *number of corrections after feedback* was the most critical predictive variable for all three clusters. However, for Clusters 1 (N = 48) and 3 (N = 16), the *no. of corrections after feedback* was negatively linked to achievement, but for Cluster 2 (N = 38), it was positively. Additionally, lower *average time on task*, *attempt times*, and *time between tasks* were associated with better predictions across all three clusters.

5.2 RQ2: Do the Students from the Identified Clusters Differ Regarding Their Displayed Learning Behaviors?

To understand whether students differ regarding their learning behavior across clusters, we investigated their displayed learning behaviors and described the average student per cluster (Table 3). Our results show that students in Cluster 1 work on average faster on tasks, spend most of their time in the system working on tasks, and have more correct answers on the first try. Overall, students in Cluster 1 performed best in the post-test ($M = 20.58, SD = 6.46$). The average Cluster 2 student takes more time for their attempts and to complete whole tasks, needs more attempts in general to solve a task field, and has fewer directly correct answers. Students in this cluster achieve average post-test scores ($M = 15.52, SD = 5.36$). Students in Cluster 3, on the other hand, work slower on their tasks, spend most of their time in the system not directly working on tasks (but rather on different pages like the task overview page), and have fewer attempts while also submitting less completely solved tasks. Those students have the lowest post-test scores ($M = 12.94, SD = 2.64$).

5.3 RQ3: Do the Students From the Identified Clusters Differ Regarding Their Prior Knowledge Levels, Motivation, or Personality?

In Clusters 1 and 2, students differed only regarding their levels of prior knowledge, with high prior knowledge in Cluster 1 and low prior knowledge in Cluster 2. Students in both clusters showed average levels of motivation and personality traits. However, students in Cluster 3 had less prior knowledge and lower motivation and conscientiousness.

In summary, Cluster 1 seems to group students with higher prior knowledge and average motivation and conscientiousness. Cluster 2 consists of students with low prior knowledge but still average motivation and conscientiousness. Cluster 3 groups students with low prior knowledge and lower motivation and conscientiousness.

6 Discussion

This study highlights the importance of considering individual differences in analyzing behavioral trace data. By identifying distinct learner clusters based on their learning behaviors' associations with performance and linking these to person-level factors, we contribute to a nuanced understanding of how digital learning tools can be optimized for diverse student populations. Educators can use these insights to tailor instructional strategies, enhance personalized learning pathways, and reduce student failure risk.

6.1 Theoretical & Practical Implications

In line with previous studies, we could show that learning behavior explains variance in achievement [7, 13]. Clustering the students based on the relevance of specific learning behavior indicators for achievement revealed three groups. Among the top ten most important predictors, there was only slight variance between groups, indicating that the same behaviors were relevant for achievement across groups. However, the strength and direction of links among learning behaviors and performance differed between groups. Notably, the levels of the most relevant variable for all three groups, *the number of corrections after feedback*, had opposite meanings for the outcome prediction. For Clusters 1 and 3, the number of corrections after feedback decreased the predicted performance, whereas for Cluster 2, it increased the predicted performance. Working faster on tasks was positively linked to higher achievement for all three clusters, possibly due to system-design aspects: The tasks in the ITS were broken down into small bits where, in most cases, only one word was needed to fill in. They were practicing opportunities for familiar concepts rather than introducing new ones. Additionally, feedback appeared only after an incorrect attempt, so lower task times might reflect fewer errors. Spending more time on tasks might, therefore, likely indicate off-task behavior rather than deep thinking. This finding highlights the advantages of using a method like SHAP that can give directional insights and emphasizes the need for domain knowledge to interpret the results helpfully [38].

Subsequent cluster comparisons revealed differences between clusters regarding students' displayed learning behaviors and their prior knowledge, motivation, and personality trait levels. Taken

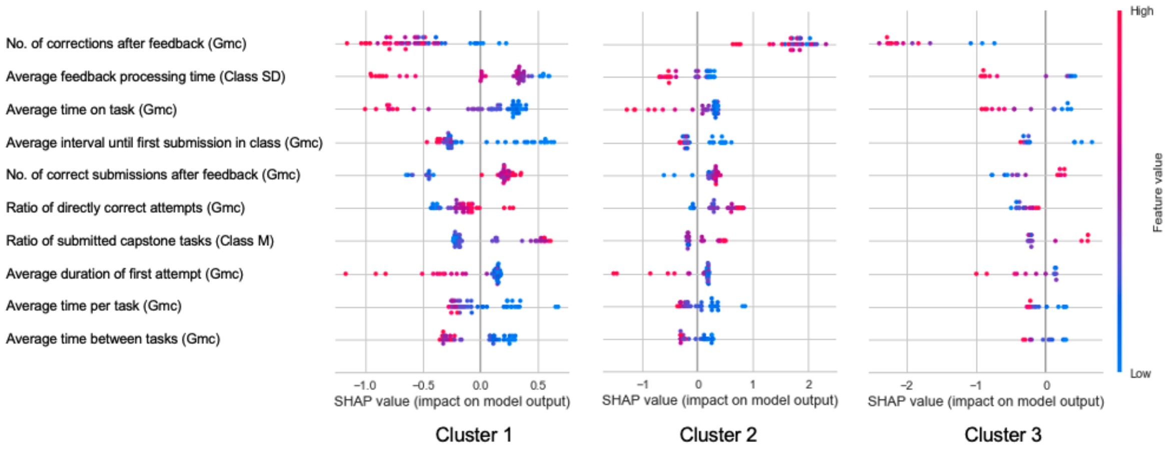

Figure 2: SHAP summary plot depicting the top ten most important learning behavior indicators for the best-performing XGBoost model in the test set split by cluster. Each point represents an individual student. The y-axis lists the ten most influential variables in the test set (the most influential at the top). The x-axis shows the SHAP values (negative values indicate a decrease and positive values indicate an increase in predicted English proficiency). The color gradient represents the variable values, ranging from low (blue) to high (red). Gmc = group mean-centered.

Table 3: Description of the average student for each cluster based on prototypical behaviors and person-level factors (using mean scores per cluster).

	Cluster 1	Cluster 2	Cluster 3
SHAP values	• most important: no. of corrections after feedback, time spent on feedback, time on task • high no. of corrections after feedback has a negative, low no. has no or a positive impact	• most important: no. of corrections after feedback, ratio of directly correct attempts, time on task • positive impact of no. of corrections after feedback	• most important: no. of corrections after feedback, time spent on feedback, time on task • no. of corrections after feedback with a generally negative impact
Learning behaviors	• low time on task • low time between tasks • more directly correct answers	• higher attempt time and time on task • higher number of attempts • fewer directly correct answers	• high time on task and time between tasks • fewer attempts • less completely solved task submitted
Person-level factors	• high prior knowledge • average motivation • average conscientiousness and openness	• low prior knowledge • average motivation • average conscientiousness and openness	• low prior knowledge • lower motivation • lower conscientiousness, average openness
English proficiency post-test score	• high to medium score	• medium score	• low score

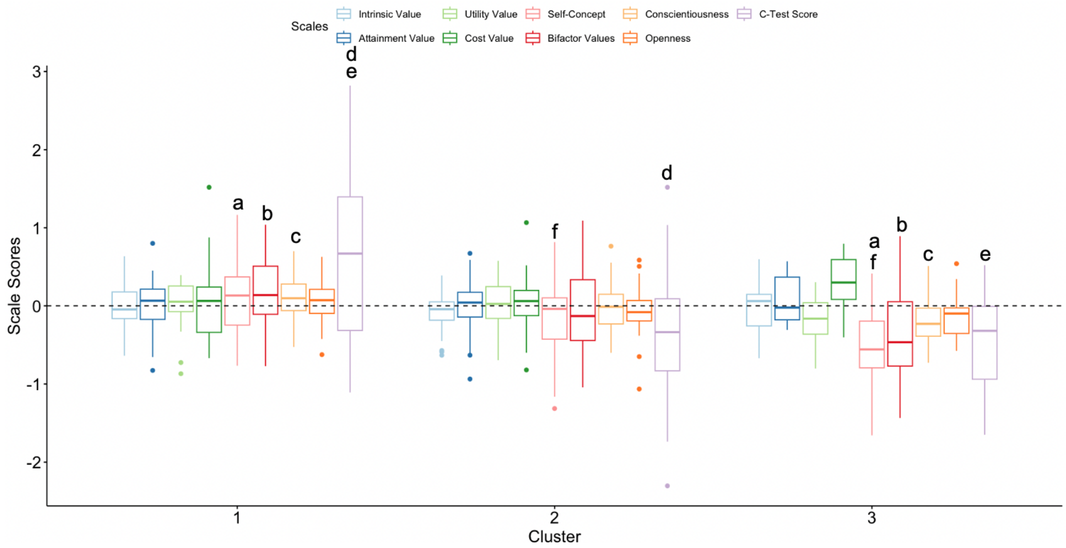

Figure 3: Box plots for prior knowledge, motivation, and personality rates by cluster. The dashed line marks the average value. The letters indicate the six statistically significant MANOVA-based pairwise comparisons between Clusters (α = 0.05): a = Self-concept for Clusters 1 and 3, b = Bifactor values for Clusters 1 and 3, c = Conscientiousness for Clusters 1 and 3, d = Prior Knowledge for Clusters 1 and 2, e = Prior Knowledge for Clusters 1 and 3, f = Self-Concept for Clusters 2 and 3.

together, the varying impact of behaviors by cluster and the corresponding individual differences provide insights into how to intervene, highlighting the importance of cluster membership for personalization. For instance, the *number of corrections after feedback* reflects differing needs: For students in Cluster 1 (high prior knowledge & average motivation), having a high number of corrections after feedback was associated with a decrease in the predicted post-test score. These students already had knowledge about the grammar constructs to be learned. Errors might result from working too fast or not being concentrated. Students in this cluster might benefit from feedback that encourages careful checking of first attempts. Additionally, future interventions could focus on challenging them to prevent complacency. In contrast, Cluster 2 students (low prior knowledge & average motivation) profited from correcting their answers after receiving feedback. This might indicate that students in this group benefited from the feedback and were motivated enough to correct their answers and learn the respective grammar construct. For Cluster 3 students (low prior knowledge & low motivation), often correcting their answers after feedback was negatively associated with their post-test score. Therefore, the behavior we see in these students might not be productive learning behavior. This is also supported by the fact that students in this group spent less time in the system working on tasks but instead, for example, on the dashboard page. These students might benefit the most from interventions targeting their motivation to learn through game-based [6] or learning-to-learn approaches targeting self-regulation or attention control skills, e.g., [8]. Within our system, the established clusters allow real-time classification of new learners after initial learning sessions. In addition, the proposed approach can be adapted for use in other systems to identify new patterns.

From a theoretical viewpoint, our findings stress the need to incorporate individual differences and their interaction effect with learning behaviors on achievement into learning analytics research. Additionally, we provide additional empirical support for relationships previously found based on self-report data with behavioral trace data indicators [37, 50]. Our proposed method to investigate such effects combined with ML, XAI, and clustering seems to lead to valuable results by better understanding differences in person-level factors when focusing on the link between learning behavior and achievement.

From a practical viewpoint, our results emphasize that behavioral interventions to increase achievement should take person-level factors into account. Our approach provides a framework for understanding and predicting achievement by accounting for behavioral differences and linking these with person-level factors. This can lead to more personalized interventions by tailoring the feedback to the students' preconditions (see [31] for an example).

6.2 Limitations and Future Work

Despite the diligent study planning, this investigation also has some limitations that one should be aware of when interpreting the results. First, there are some constraints related to the system itself, such as the fact that we had no control over and no knowledge of learning behavior that happened outside the system. Additionally, as is true for all ITS, the interpretation of our findings is specific to the context in which the data is generated [38], meaning they are not necessarily transferable to other digital learning environments. A helpful next step would be to investigate the applicability of the results in another task cycle (within sample stability). Then, investigating clusters in other samples would help establish out-of-sample stability. This limitation to the context, however, bears great potential for future research to extend our findings across educational settings. Yet, while our study was conducted within a specific platform, the methodological framework is adaptable to various contexts by adjusting the behavioral metrics.

Second, our study's advantage of investigating learning within classrooms in an ecologically valid setting also causes one disadvantage. Some teachers did not use the system with their classes as expected, i.e., students have occasionally worked in groups while logged in to only one student account during class, which blurs the measured learning behavior. Additionally, the tasks students worked on depended highly on the tasks the teachers asked them to do. Further, some system features were only used by a few students, such as visits to grammar help pages.

Third, our sample is typical for academic-track schools regarding their socioeconomic status (SES) and only consists of one age group. This might limit the generalizability to other age groups and school types with differing SES.

Fourth, in terms of how learning behavior is modeled, we did not conduct temporal analyses, which might inform how learning takes place holistically. Conducting analyses that consider the temporal aspects of learning could be a fruitful next step [29]. Additionally, k-means clustering can only provide results for a specific sample, and results might depend on the initial setup. By bootstrapping parts of our analyses, we aimed to ensure stable results; however, it might make the results even more stable by bootstrapping the clustering procedure with multiple random seeds.

7 Conclusion

In summary, this study demonstrates the utility of combining ML, XAI, and clustering to model the moderating effect of person-level indicators on the link between learning behaviors and achievement. This approach generated a better understanding of learning processes as they took place in an ecologically valid setting and paved the way for genuinely personalized interventions tailored to individual prerequisites. In addition, it has enabled researchers to investigate the effects that have previously primarily been examined based on self-report data with behavioral trace data and ML.

Acknowledgments

Hannah Deininger is a doctoral student at the LEAD Graduate School & Research Network funded by the Ministry of Science, Research and the Arts of Baden-Württemberg. This research was supported by the German Federal Ministry of Research and Education (BMBF; 01JD1905A and 01JD1905B). Part of this research was developed in a cooperation with the aim Heilbronn financially supported by the Dieter Schwarz Foundation. ChatGPT was used for language editing.

References

[1] Amjed Abu Saa, Mostafa Al-Emran, and Khaled Shaalan. 2019. Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques. *Technology, Knowledge and Learning* 24, 4 (Dec. 2019), 567–598. https://doi.org/10.1007/s10758-019-09408-7

[2] Herman Aguinis, Ryan K. Gottfredson, and Harry Joo. 2013. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods* 16, 2 (April 2013), 270–301. https://doi.org/10.1177/1094428112470848

[3] Miguel Alvarez-Garcia, Mar Arenas-Parra, and Raquel Ibar-Alonso. 2024. Uncovering student profiles. An explainable cluster analysis approach to PISA 2022. *Computers & Education* 223 (2024), 105166.

[4] Cara J. Arizmendi, Matthew L. Bernacki, Mladen Raković, Robert D. Plumley, Christopher J. Urban, A. T. Panter, Jeffrey A. Greene, and Kathleen M. Gates. 2023. Predicting Student Outcomes Using Digital Logs of Learning Behaviors: Review, Current Standards, and Suggestions for Future Work. *Behavior Research Methods* 55, 6 (2023), 3026–3054. https://doi.org/10.3758/s13428-022-01939-9

[5] Mariette Awad, Rahul Khanna, Mariette Awad, and Rahul Khanna. 2015. *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*. Springer, New York, Chapter Support vector regression, 67–80.

[6] Nathalie Barz, Manuela Benick, Laura Dörrenbächer-Ulrich, and Franziska Perels. 2024. The Effect of Digital Game-Based Learning Interventions on Cognitive, Metacognitive, and Affective-Motivational Learning Outcomes in School: A Meta-Analysis. *Review of Educational Research* 94, 2 (2024), 193–227. https://doi.org/10.3102/00346543231167795

[7] Matthew L. Bernacki, Michelle M. Chavez, and P. Merlin Uesbeck. 2020. Predicting Achievement and Providing Support before STEM Majors Begin to Fail. *Computers & Education* 158 (Dec. 2020), 103999. https://doi.org/10.1016/j.compedu.2020.103999

[8] Matthew L Bernacki, Lucie Vosicka, and Jenifer C Utz. 2020. Can a brief, digital skill training intervention help undergraduates "learn to learn" and improve their STEM achievement? *Journal of Educational Psychology* 112, 4 (2020), 765.

[9] Carolyn Blume, Lisa Middelanis, and Torben Schmidt. 2024. Where Tasks, Technology, and Textbooks Meet: An Exploratory Analysis of English Language Teachers' Perceived Affordances of an Intelligent Language Tutoring System. In *Textbook 4.0–From Paper-Based Textbooks with Digital Components to Interactive Teaching and Learning Environments*. Peter Lang International Academic Publishers, Bern, 125–161. https://library.oapen.org/handle/20.500.12657/94835

[10] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2022. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems* 35, 6 (2022), 7499–7519.

[11] Eva Bosch, Eva Seifried, and Birgit Spinath. 2021. What Successful Students Do: Evidence-based Learning Activities Matter for Students' Performance in Higher Education beyond Prior Knowledge, Motivation, and Prior Achievement. *Learning and Individual Differences* 91 (Oct. 2021), 102056. https://doi.org/10.1016/j.lindif.2021.102056

[12] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[13] Tianqi Chen and C. Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[14] Leona Colling, Ines Pieronczyk, Cora Parrisius, Heiko Holz, Stephen Bodnar, Florian Nuxoll, and Detmar Meurers. 2024. Towards Task-Oriented ICALL: A Criterion-Referenced Learner Dashboard Organising Digital Practice. In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024)*, Vol. 1. SciTePress, Angers, France, 668–679. https://doi.org/10.5220/0012753000003693

[15] P.G. De Barba, G.E. Kennedy, and M.D. Ainley. 2016. The Role of Students' Motivation and Participation in Predicting Performance in a MOOC: Motivation and Participation in MOOCs. *Journal of Computer Assisted Learning* 32, 3 (June 2016), 218–231. https://doi.org/10.1111/jcal.12130

[16] Hannah Deininger, Rosa Lavelle-Hill, Cora Parrisius, Ines Pieronczyk, Leona Colling, Detmar Meurers, Ulrich Trautwein, Benjamin Nagengast, and Gjergji Kasneci. 2023. Can You Solve This on the First Try? – Understanding Exercise Field Performance in an Intelligent Tutoring System. In *Artificial Intelligence in Education*, Ning Wang, Genaro Rebolledo-Mendez, Noboru Matsuda, Olga C. Santos, and Vania Dimitrova (Eds.). Vol. 13916. Springer Nature Switzerland, Cham, 565–576. https://doi.org/10.1007/978-3-031-36272-9_46

[17] Hannah Deininger, Ines Pieronczyk, Cora Parrisius, Robert D Plumley, Detmar Meurers, Gjergji Kasneci, Benjamin Nagengast, Ulrich Trautwein, Jeffrey A Greene, and Matthew L Bernacki. 2024. Using theory-informed learning analytics to understand how homework behavior predicts achievement. *Journal of Educational Psychology* (2024), Advance online publication. https://doi.org/10.1037/edu0000906

[18] Statistisches Bundesamt (Destatis). 2020. Bildungsstand der Bevölkerung - Ergebnisse des Mikrozensus 2019 [Educational level of the population - Results of the microcensus 2019]. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Bildungsstand/Publikationen/Downloads-Bildungsstand/bildungsstand-bevoelkerung-5210002197004.pdf?__blob=publicationFile.

[19] Jacquelynne S. Eccles and Allan Wigfield. 2020. From Expectancy-Value Theory to Situated Expectancy-Value Theory: A Developmental, Social Cognitive, and Sociocultural Perspective on Motivation. *Contemporary Educational Psychology* 61 (2020), 101859–101859.

[20] Anastasia Efklides. 2011. Interactions of Metacognition With Motivation and Affect in Self-Regulated Learning: The MASRL Model. *Educational Psychologist*

46, 1 (Jan. 2011), 6–25. https://doi.org/10.1080/00461520.2011.538645

[21] Anastasia Efklides and Bennett L Schwartz. 2024. Revisiting the Metacognitive and Affective Model of Self-Regulated Learning: Origins, Development, and Future Directions. *Educational Psychology Review* 36, 2 (2024), 61.

[22] Ed Fincham, Alexander Whitelock-Wainwright, Vitomir Kovanović, Srećko Joksimović, Jan-Paul Van Staalduinen, and Dragan Gašević. 2019. Counting Clicks Is Not Enough: Validating a Theorized Model of Engagement in Learning Analytics. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. ACM, Tempe AZ USA, 501–510. https://doi.org/10.1145/3303772.3303775

[23] Barbara Flunger, Ulrich Trautwein, Benjamin Nagengast, Oliver Lüdtke, Alois Niggli, and Inge Schnyder. 2017. A Person-Centered Approach to Homework Behavior: Students' Characteristics Predict Their Homework Learning Type. *Contemporary Educational Psychology* 48 (Jan. 2017), 1–15. https://doi.org/10.1016/j.cedpsych.2016.07.002

[24] Hanna Gaspard, Isabelle Häfner, Cora Parrisius, Ulrich Trautwein, and Benjamin Nagengast. 2017. Assessing Task Values in Five Subjects during Secondary School: Measurement Structure and Mean Level Differences across Grade Level, Gender, and Academic Subject. *Contemporary Educational Psychology* 48 (2017), 67–84. https://doi.org/10.1016/j.cedpsych.2016.09.003

[25] Arto Hellas, Petri Ihantola, Andrew Petersen, Vangel V. Ajanovski, Mirela Gutica, Timo Hynninen, Antti Knutas, Juho Leinonen, Chris Messom, and Soohyun Nam Liao. 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*. ACM, Larnaca Cyprus, 175–199. https://doi.org/10.1145/3293881.3295783

[26] Maria Hendriks, Hans Luyten, Jaap Scheerens, and Peter Sleegers. 2014. Meta-Analyses. In *Effectiveness of Time Investments in Education: Insights from a review and meta-analysis*. Springer, Heidelberg, 55–142.

[27] Bronson Hui, Björn Rudzewitz, and Detmar Meurers. 2023. Learning Processes in Interactive CALL Systems: Linking Automatic Feedback, System Logs, and Learning Outcomes. *Language Learning & Technology* 27, 1 (2023), 1–23. https://doi.org/10.125/73527

[28] Paul Hur, HaeJin Lee, Suma Bhat, and Nigel Bosch. 2022. Using Machine Learning Explainability Methods to Personalize Interventions for Students. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM)*. ERIC, Durham, United Kingdom, 438–446.

[29] Simon Knight, Alyssa Friend Wise, and Bodong Chen. 2017. Time for change: Why learning analytics needs temporal analysis. *Journal of Learning Analytics* 4, 3 (2017), 7–17.

[30] James A Kulik and John D Fletcher. 2016. Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research* 86, 1 (2016), 42–78.

[31] Amit Kumar and N. J. Ahuja. 2020. An Adaptive Framework of Learner Model Using Learner Characteristics for Intelligent Tutoring Systems. In *Intelligent Communication, Control and Devices*, Sushabhan Choudhury, Ranjan Mishra, Raj Gaurav Mishra, and Adesh Kumar (Eds.). Springer Singapore, Singapore, 425–433. https://doi.org/10.1007/978-981-13-8618-3_45

[32] Luis Earving Lee Hernández, José Antonio Castán-Rocha, Salvador Ibarra-Martínez, Jésus David Terán-Villanueva, Mayra Guadalupe Treviño-Berrones, and Julio Laria-Menchaca. 2023. Cluster Analysis Using K-Means in School Dropout. In *Data Analytics and Computational Intelligence: Novel Models, Algorithms and Applications*, Gilberto Rivera, Laura Cruz-Reyes, Bernabé Dorronsoro, and Alejandro Rosete (Eds.). Vol. 132. Springer Nature Switzerland, Cham, 3–16. https://doi.org/10.1007/978-3-031-38325-0_1

[33] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017), 1–10.

[34] Wannisa Matcha, Dragan Gašević, Jelena Jovanović, Nora'ayu Ahmad Uzir, Chris W Oliver, Andrew Murray, and Danijela Gasevic. 2020. Analytics of Learning Strategies: The Association with the Personality Traits. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. ACM, Frankfurt Germany, 151–160. https://doi.org/10.1145/3375462.3375534

[35] Edward McAuley, Terry Duncan, and Vance V. Tammen. 1989. Psychometric Properties of the Intrinsic Motivation Inventory in a Competitive Sport Setting: A Confirmatory Factor Analysis. *Research Quarterly for Exercise and Sport* 60, 1 (March 1989), 48–58. https://doi.org/10.1080/02701367.1989.10607413

[36] Christoph Molnar. 2020. *Interpretable machine learning*. Leanpub.

[37] Hoi Kwan Ning and Kevin Downing. 2012. Influence of student learning experience on academic performance: The mediator and moderator effects of self-regulation and motivation. *British Educational Research Journal* 38, 2 (2012), 219–237.

[38] Larian M Nkomo, Ben K Daniel, and Russell J Butson. 2021. Synthesis of student engagement with digital technologies: a systematic review of the literature. *International Journal of Educational Technology in Higher Education* 18 (2021), 1–26.

[39] Cora Parrisius, Ines Pieronczyk, Carolyn Blume, Katharina Wendebourg, Diana Pili-Moss, Mirjam Assmann, Sabine Beilharz, Stephen Bodnar, Leona Colling, Heiko Holz, Lisa Middelanis, Florian Nuxoll, Julia Schmidt-Peterson, Detmar Meurers, Benjamin Nagengast, Torben Schmidt, and Ulrich Trautwein. 2022. Using an Intelligent Tutoring System within a Task-Based Learning Approach in English as a Foreign Language Classes to Foster Motivation and Learning Outcome (Interact4School): Pre-registration of the Study Design. *PsychArchives* (2022), 1–74. https://doi.org/10.23668/psycharchives.5366

[40] Cora Parrisius, Katharina Wendebourg, Sven Rieger, Carolyn Blume, Diana Pili-Moss, Leona Colling, Ines Pieronczyk, Heiko Holz, Stephen Bodnar, Ines Loll, Thorben Schmidt, Ulrich Trautwein, Detmar Meurers, and Benjamin Nagengast. unpublished manuscript. Effective features of feedback in an intelligent language tutoring system. (unpublished manuscript). Institute for Educational Research Methods, Karlsruhe University of Education.

[41] Rachel Part, Harsha N. Perera, Kyle Mefferd, and Chyna J. Miller. 2023. Decomposing Trait and State Variability in General and Specific Subjective Task Value Beliefs. *Contemporary Educational Psychology* 72 (Jan. 2023), 102112. https://doi.org/10.1016/j.cedpsych.2022.102112

[42] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[43] R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[44] Ulhich Raatz and Christine Klein-Braley. 1981. *The C-Test–A Modification of the Cloze Procedure*. ERIC, United States.

[45] Beatrice Rammstedt and Oliver P. John. 2005. Kurzversion Des Big Five Inventory (BFI-K):. *Diagnostica* 51, 4 (Oct. 2005), 195–206. https://doi.org/10.1026/0012-1924.51.4.195

[46] Lior Rokach and Oded Maimon. 2005. Decision Trees. In *Data Mining and Knowledge Discovery Handbook*, Oded Maimon and Lior Rokach (Eds.). Springer US, Boston, MA, 165–192. https://doi.org/10.1007/0-387-25465-X_9

[47] Cristóbal Romero, José Raúl Romero, and Sebastián Ventura. 2014. A Survey on Pre-Processing Educational Data. In *Educational Data Mining*, Alejandro Peña-Ayala (Ed.). Vol. 524. Springer International Publishing, Cham, 29–64. https://doi.org/10.1007/978-3-319-02738-8_2

[48] Cristobal Romero and Sebastian Ventura. 2020. Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery* 10, 3 (2020), 1–21.

[49] Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, and Detmar Meurers. 2017. Developing a web-based workbook for English supporting the interaction of students and teachers. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*. Linköping Electronic Conference Proceedings, Sweden, 36–46. https://aclanthology.org/W17-0305

[50] Michael Schneider and Franzis Preckel. 2017. Variables Associated with Achievement in Higher Education: A Systematic Review of Meta-Analyses. *Psychological Bulletin* 143, 6 (2017), 565–600. https://doi.org/10.1037/bul0000098

[51] Anan Schütt, Tobias Huber, Jauwairia Nasir, Cristina Conati, and Elisabeth André. 2024. Does Difficulty Even Matter? Investigating Difficulty Adjustment and Practice Behavior in an Open-Ended Learning Task. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. ACM, Kyoto Japan, 253–262. https://doi.org/10.1145/3636555.3636876

[52] scikit-learn. 2024. 2.3. Clustering. https://scikit-learn.org/1.5/modules/clustering.html.

[53] Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster Quality Analysis Using Silhouette Score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, United States, 747–748. https://doi.org/10.1109/DSAA49011.2020.00096

[54] Abhishek Sheetal, Zhou Jiang, and Lee Di Milia. 2023. Using machine learning to analyze longitudinal data: A tutorial guide and best-practice recommendations for social science researchers. *Applied Psychology* 72, 3 (2023), 1339–1364. https://doi.org/10.1111/apps.12435

[55] Michelle Taub and Roger Azevedo. 2019. How Does Prior Knowledge Influence Eye Fixations and Sequences of Cognitive and Metacognitive SRL Processes during Learning with an Intelligent Tutoring System? *International Journal of Artificial Intelligence in Education* 29, 1 (March 2019), 1–28. https://doi.org/10.1007/s40593-018-0165-4

[56] Michelle Taub, Allison M. Banzon, Tom Zhang, and Zhongzhou Chen. 2022. Tracking Changes in Students' Online Self-Regulated Learning Behaviors and Achievement Goals Using Trace Clustering and Process Mining. *Frontiers in Psychology* 13 (March 2022), 813514. https://doi.org/10.3389/fpsyg.2022.813514

[57] Dirk Tempelaar, Bart Rienties, and Quan Nguyen. 2021. The Contribution of Dispositional Learning Analytics to Precision Education. *Educational Technology & Society* 24, 1 (2021), 109–121.

[58] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.

[59] Jian Wang and Yuanyuan Zhang. 2019. Clustering study of student groups based on analysis of online learning behavior. In *Proceedings of the 2019 International Conference on Modern Educational Technology*. Association for Computing Machinery, New York, NY, USA, 115–119. https://doi.org/10.1145/3341042.3341065

[60] Rand R. Wilcox. 2011. *Introduction to Robust Estimation and Hypothesis Testing*. Academic press, Amsterdam.

[61] Alyssa Friend Wise and David Williamson Shaffer. 2015. Why Theory Matters More than Ever in the Age of Big Data. *Journal of Learning Analytics* 2, 2 (Dec. 2015), 5–13. https://doi.org/10.18608/jla.2015.22.2

[62] Jianzhong Xu. 2022. More than Minutes: A Person-Centered Approach to Homework Time, Homework Time Management, and Homework Procrastination. *Contemporary Educational Psychology* 70 (July 2022), 102087. https://doi.org/10.1016/j.cedpsych.2022.102087