To What Extent Can Early Keystroke Patterns Predict Knowledge Gain in Mathematics

Oscar Blessed Deho
Centre for Change and Complexity in Learning (C3L)
Adelaide University
Australia
oscar.deho@adelaide.edu.au

Ryan S. Baker
Centre for Change and Complexity in Learning (C3L)
Adelaide University
Australia
ryan.baker@adelaide.edu.au

Jiayi Zhang
Penn Center for Learning Analytics
University of Pennsylvania
United States
joycez@upenn.edu

Digory Smith
Eedi
United Kingdom
digory.smith@eedi.com

Simon Woodhead
Eedi
United Kingdom
simon.woodhead@eedi.com

ABSTRACT

Early identification of students who may struggle to learn from practice could enable timely instructional support. This study investigates whether keystroke dynamics during initial problem-solving attempts can signal subsequent knowledge gain in mathematics. We analyzed keystroke patterns from students' first two problems following instruction on a skill, examining associations with knowledge gain operationalized through Bayesian Knowledge Tracing. Students who ultimately learned more exhibited systematically different early behaviors: slower, more variable typing with greater revision activity. Effect sizes ranged from small to moderate (Cliff's $\delta = 0.15$–0.38), with timing features showing strongest associations. Machine learning models achieved modest but above-chance predictive accuracy (regression: RMSE = 0.36, $r = 0.42$; classification: AUC-ROC = 0.68, balanced accuracy = 0.67). While individual keystroke features explained limited variance (3.6–11.6%), results suggest that fine-grained behavioral traces from students' earliest practice attempts may provide meaningful early signals about learning trajectories, potentially enabling proactive instructional adaptation before patterns of persistent struggle emerge.

Keywords

Keystroke dynamics, Knowledge gain, Early prediction

1. INTRODUCTION

Intelligent tutoring systems (ITSs) have demonstrated the potential to significantly improve student learning outcomes [14]. By personalizing instruction and feedback, these systems can help learners acquire new knowledge and skills more effectively than traditional one-size-fits-all instruction. However, an important goal for such systems is not merely to optimize immediate performance (e.g., getting questions correct), but to ensure that students are actually learning—making durable knowledge gains [13]. In practice, a correct answer does not always equate to genuine understanding. For instance, some learners might arrive at correct answers by rote application of superficial strategies, such as relying heavily on hints, pattern matching, or completing problems through trial and error, without developing a deeper understanding of the underlying concepts or procedures. Such behavior may lead to high performance in the short term while yielding smaller gains in robust, long-term learning. This contrast between performance and learning highlights the need for adaptive platforms that focus on cultivating knowledge growth (e.g., long-term retention and transfer; [13]), rather than merely supporting immediate problem correctness.

A related challenge for intelligent tutoring systems concerns when different kinds of instructional decisions should be made. Most ITSs already provide rapid, fine-grained support at the inner-loop level—for example, responding immediately to an incorrect step with hints or feedback [17]. This form of in-the-moment guidance is often effective, but it is not always sufficient. Some students fail to make substantive progress despite repeated inner-loop assistance, resulting in wheel-spinning, where they continue working over a substantial time without moving closer to mastery [4]. Addressing such situations can require intervention by a teacher or outer-loop decisions by the learning environment [17], such as providing support with prerequisite skills. In most existing systems, however, these types of interventions only occur when the instructor realizes that a student has been struggling for quite some time. Although prior work has explored early prediction of eventual wheel-spinning [15, 5, 18], these predictions still rely on several problems of student interaction. This raises a natural question: can the need for deeper support be inferred even earlier—during a student's initial encounters with a skill? If so, an ITS could begin adapting instruction not after prolonged difficulty, but from the very beginning of work on the skill.

Early prediction of learning outcomes requires access to in-

formation that reflects how students engage with tasks as learning unfolds. Beyond final answers, fine-grained interaction traces—such as timing, attempts, and patterns of interaction—can reveal how a student is approaching a problem in real time [12]. Keystroke-level information provides even finer-grained detail, capturing the exact sequence and timing of keys pressed as a student constructs a response. This level of detail has been argued to shed light on the cognitive and metacognitive processes involved in problem solving [9]. For example, long pauses before initiating a response may reflect planning or recall, while rapid sequences of edits and deletions may indicate uncertainty or guess-and-revise strategies. In this sense, how a student types may provide insight into hesitation, confidence, and strategy use—factors that are closely related to learning [1, 16].

Building on this motivation, prior research in educational data mining has explored early prediction of learning outcomes using students' initial interactions [4, 6, 15, 5, 18]. These studies draw on data such as keystrokes, clickstream logs, and tutor transaction sequences to anticipate short- and long-term outcomes. For instance, Beck and Gong's work on wheel-spinning examined a student's first few attempts on a new skill to identify students unlikely to attain mastery, flagging those at risk of continuous failure and prompting possible changes in learning support. Other studies have focused on course-level or test outcomes. Casey [6] developed classifiers to detect at-risk programming students early in the semester; the inclusion of keystroke features improved the model's ability to predict which students would not pass the course. Conijn et al. [7] examined keystroke logs from the beginnings of writing assignments to forecast final essay quality. While their predictive models only modestly outperformed baselines, they demonstrated the viability of using typing process data to identify students who might need writing support before the draft is completed. Researchers have also shown that even a short window of student activity can yield insight into far-future outcomes. Gao et al. [10] found that data from the first 2–5 hours of student usage of an educational game and an ITS provided a valuable signal about students' performance on end-of-year assessments months later. In their study, a machine learning model using only the initial few hours of log data was able to predict year-end test scores with notable accuracy, indicating that early behaviors encapsulate information about eventual learning results. Together, this body of literature illustrates the promise of early prediction: from fine-grained keystroke patterns to aggregated usage metrics, a range of early indicators have been leveraged to forecast outcomes such as course failure, essay scores, and standardized test performance.

While prior studies demonstrate that early interaction data, including keystrokes, can be informative for predicting later performance or task outcomes, they offer limited insight into whether such early behaviors relate to subsequent knowledge gain during skill practice. Work incorporating keystroke features has examined outcomes such as essay quality, programming performance, or mastery likelihood, rather than learning improvement across a sequence of problems. Consequently, it remains unclear whether keystroke behaviors observed during students' initial encounters with a new skill are related to how much they ultimately learn through continued practice. This gap matters because proactive support requires identifying not only students who struggle to answer correctly, but also those who make limited learning gains even when their early answers are correct.

To fill this gap, we address two research questions. First, how are students' early keystroke patterns during their initial encounters with a skill associated with subsequent knowledge gain? Second, to what extent can early keystroke patterns predict students' eventual knowledge gain on that skill? To investigate these questions, we analyze students' keystroke patterns from the first two problems of a skill and relate them to subsequent knowledge gain. Using Bayesian Knowledge Tracing (BKT), we operationalize knowledge gain as the difference between the student's BKT-estimated mastery level at the start and end of practice on a given skill. We employ both statistical analyses and machine learning approaches. We examine group differences in early keystroke features between high-gain and low-gain students, then train predictive models to assess whether early keystroke features contain a meaningful signal about students' eventual knowledge gain. One modeling task estimates the amount of learning a student will achieve, while a complementary classification task evaluates whether these early behaviors can distinguish students who ultimately show higher versus lower knowledge gains.

2. METHODS

2.1 Data

2.1.1 Source and Context

This study used de-identified data from an online mathematics learning platform with adaptive practice. Each topic begins with a short diagnostic assessment, after which students receive targeted instruction through worked examples and videos before entering the worksheet practice phase. We focus on this worksheet phase, where students independently solve problems using open-ended typed responses that generate keystroke sequences and timing information. All personally identifiable information was removed prior to data access. The final dataset comprised 19,155 fully matched student–question transactions, each representing a single practice problem.

2.1.2 Correctness Labelling

The worksheet data did not include correctness labels. To generate these at scale, we developed and validated an automated correctness labeler using GPT-4o, prompted to assess whether a student's response was mathematically equivalent to an expert reference answer under explicit equivalence rules (e.g., commutativity, numerical approximation, and formatting variations). Ground truth was established via double human annotation of 200 responses, yielding high inter-rater reliability ($\kappa = 0.95$); disagreements were resolved through social moderation. The automated labeler closely matched human judgments ($\kappa = 0.95$, precision = 0.97, recall = 0.95) and was subsequently used to label all 19,155 responses.

2.1.3 Keystroke Data and Feature Engineering

Each worksheet question answered generated a sequence of keystrokes and associated metadata, including information

Table 1: Example of selected columns from a single row of the raw data (← = left-arrow cursor movement; BKSP = Backspace).

Question	Input	Keystroke	Reference	Time(s)
2548 − 362	2186	← 218y BKSP 6	2186	167

Table 2: Features and their definitions.

Feature	Definition
Keystroke edit distance	Levenshtein distance between the full raw keystroke sequence and the final submitted answer.
Seconds per keystroke	Total time spent on the response divided by the total number of keystrokes.
Keystroke entropy	Shannon entropy of the keystroke sequence.
Correction density	Number of correction keys (Backspace, Delete) divided by log-transformed total time.
Correction proportion	Number of correction keys divided by total keystrokes.
Keystroke efficiency ratio	Final answer length divided by total keystrokes.
Average seconds between keys	Mean inter-keystroke time (seconds) for a question attempt.
Variation in seconds between keys	Standard deviation of inter-keystroke times (seconds) for a question attempt.

about time spent, editing actions, and overall input behavior. Table 1 illustrates an example of a single row of data showing the information captured for a student's response to a worksheet question. From these data, a set of features was derived to capture temporal, corrective, and compositional aspects of students' typing activity. Table 2 presents a detailed summary of all engineered features.

2.2 Analysis

2.2.1 Deriving Knowledge Gain from BKT

To quantify learning within each skill, we applied Bayesian Knowledge Tracing (BKT) [8] to students' sequential response data. BKT models student learning as a dynamic process within a Hidden Markov Model framework, where each student transitions between two latent states: not mastered and mastered across successive items answered within the same skill. While prior work has noted limitations in BKT's accuracy at predicting subsequent performance [11], we adopted BKT not for predictive accuracy but as a relative indicator of student mastery based on prior performance. For this purpose, it is useful that BKT provides a situation-independent estimate of mastery that does not incorporate item difficulty or information from other skills, unlike more recent knowledge tracing models. BKT's ability to account for guessing and slipping further supports its use for mastery estimation in this context. BKT parameters were fit using a Brute Force/Grid Search algorithm [3], with a minimum value of 0.01 applied to all probabilities and a maximum value of 0.3 imposed on guess and slip parameters to prevent model degeneracy [2]. The resulting parameter estimates were used to compute predicted probabilities of mastery for each student after each question. For each student–skill pair, we extracted the probability of mastery before any question was answered, $P(L_0)$, and after the final question, $P(L_T)$; knowledge gain was then defined as $P(L_T) - P(L_0)$. Positive values indicate increased mastery across practice, while negative values reflect performance below the initial estimate. To focus on skills where meaningful learning occurred, we retained only skills with an average gain greater than 0.1. These knowledge gain values served as the target variable for all subsequent analyses.

2.2.2 Data Preparation for Analysis

After retaining only skills with average knowledge gain above 0.1, the dataset was reduced from 19,155 to 1,212 transactions. To capture early engagement, predictors were computed from the first two questions per skill by averaging numeric features across those attempts. This two-question window balances early identification with robustness, avoiding reliance on a single initial response that may reflect task familiarization rather than learning difficulty. Applying this restriction yielded a final dataset of 232 student–skill records from 190 (unique) students across seven skills. The dataset includes all keystroke features in Table 2 and a continuous BKT-based knowledge gain label, which ranged from -0.41 to 0.99 ($M = 0.19$, $SD = 0.40$). For classification, 53% of observations fell above the median knowledge gain and 47% below, which we treat as proxies for successful and unsuccessful learning.

2.2.3 Exploratory Analysis

To investigate the relationship between early keystroke features and subsequent knowledge gain, we conducted a set of exploratory statistical analyses to characterize the nature and direction of these relationships. We tested for significant differences between students categorized as successful learners (knowledge gain at or above the skill-specific median) and unsuccessful learners (knowledge gain below the skill-specific median) using the Mann–Whitney U test. p-values were corrected for multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) procedure, and effect sizes were computed using Cliff's Delta. Following this group comparison, continuous associations between each keystroke feature and numerical knowledge gain were assessed using Spearman's rank correlation. To further explore these relationships, the data were divided into four quartiles of knowledge gain (Q1–Q4), allowing inspection of median keystroke feature values within each quartile and revealing progressive trends as knowledge gain increased. These analyses characterize individual feature–outcome relationships, while Section 2.2.4 evaluates their combined predictive power.

2.2.4 Predictive Modelling

To evaluate the predictive power of students' early keystroke patterns, we conducted two predictive modeling tasks: a regression task to predict the magnitude of knowledge gain and a classification task to identify whether learning outcomes were successful. All models were evaluated using five-fold cross-validation, stratified at the student level to ensure that data from a given student did not appear in both training and test sets within a fold. This approach prevents data leakage and supports generalization to un-

Table 3: Comparison of keystroke features between successful (S) and unsuccessful (U) learning groups via Mann-Whitney U test (group alpha = 0.05)

Feature	S_{Mdn}	U_{Mdn}	δ	p-value
average seconds between keys	1.15	0.62	**0.38**	< 0.001
variation in seconds between keys	1.02	0.40	**0.34**	< 0.001
seconds per keystroke	5.03	2.78	**0.33**	< 0.001
keystroke edit distance	1.50	1.00	0.27	0.001
correction density	0.00	0.00	0.22	0.001
correction proportion	0.00	0.00	0.20	0.002
keystroke efficiency ratio	0.71	0.80	-0.19	0.015
keystroke entropy	2.00	1.91	0.15	0.042

Table 4: Spearman correlations (ρ) between keystroke features and knowledge gain (group alpha = 0.05)

Feature	ρ	p-value
average seconds between keys	**0.34**	< 0.001
variation in seconds between keys	0.30	< 0.001
seconds per keystroke	0.30	< 0.001
correction density	0.25	< 0.001
keystroke edit distance	0.24	< 0.001
correction proportion	0.23	< 0.001
keystroke entropy	0.19	0.005
keystroke efficiency ratio	-0.15	0.020

Table 5: Median feature values across quartiles of knowledge gain. Q1= Lowest, Q4= Highest

Feature	Q1	Q2	Q3	Q4
average seconds between keys	0.58	0.66	1.08	1.16
keystroke edit distance	1.00	1.50	1.00	2.00
seconds per keystroke	2.78	2.68	4.87	6.16
keystroke entropy	1.91	1.79	1.95	2.20
correction density	0.00	0.00	0.00	0.13
correction proportion	0.00	0.00	0.00	0.04
variation in seconds between keys	0.39	0.50	0.85	1.23
keystroke efficiency ratio	0.80	0.73	0.76	0.62

seen students. Within each fold, hyperparameter optimization was performed on the training set to identify the best-performing configuration for each model.

The regression task aimed to predict the continuous value of a student's knowledge gain from early keystroke features, serving an exploratory purpose by modeling the relationship between keystroke behavior and learning magnitude. The evaluated models included Random Forest Regressor, Extra Trees Regressor, Decision Tree Regressor, Linear Regression, Ridge Regression, and M5Prime. Model performance was evaluated using linear correlation and root mean squared error (RMSE), capturing both the strength of association between predicted and observed knowledge gain and the magnitude of prediction error. All reported results were averaged across cross-validation folds. The classification task predicted whether a student's knowledge gain was above or below the median for a given skill. This task evaluates the practical diagnostic utility of early keystroke features for distinguishing between high-gain and low-gain learners, with implications for targeted intervention. The target variable was binarized relative to the median per skill to account for differences in difficulty across skills. Models evaluated included Logistic Regression, Decision Tree, Random Forest, RIPPER, and K*. Performance was measured using the Area Under the Receiver Operating Characteristic curve (AUC-ROC), the Area Under the Precision-Recall curve (AUC-PR), F1-score, standard accuracy (ACC), and balanced accuracy (BACC).

3. RESULTS

3.1 How is Early Keystroke Patterns associated with Knowledge Gain?

To understand the relationship between early keystroke features and knowledge gain, we began by comparing students classified as successful learners with their unsuccessful peers. The Mann–Whitney U test (with Benjamini–Hochberg correction) identified significant differences for all eight features (Table 3). Effect sizes (Cliff's Delta) ranged from 0.15 to 0.38, indicating small to moderate group differences. The largest effects were found for *average seconds between keys*, *variation in seconds between keys*, and *seconds per keystroke*. For these latency-based features, the effect size range ($\delta = 0.33$–0.38) corresponds to approximately a 67% to 69% probability—based on the standard Cliff's Delta derivation—that a randomly selected student from the successful learning group would exhibit a higher feature value than one from the unsuccessful group. Again, as shown in Table 3, students with higher knowledge gains consistently showed higher median values for features related to latency, variability, and editing (e.g., *average seconds between keys*, *keystroke edit distance*), suggesting that slower, more variable, and more corrective early typing behaviors are associated with successful learning. Corresponding findings were obtained when examining continuous associations. Spearman's rank correlations were significant and positive for all features, with coefficients (ρ) ranging from 0.19 to 0.34 (Table 4). This indicates that individual keystroke features explained between 3.6% and 11.6% (ρ^2) of the variance in knowledge gain ranks, with latency-related features again showing the strongest associations. This general upward trend was also visible in the quantile analysis (Table 5), which showed the median values for features such as *average seconds between keys* increasing from the lowest quartile (Q1) to the highest quartile (Q4). These statistical analyses indicate that significant relationships exist between early keystroke features and knowledge gain, motivating the subsequent investigation into their combined predictive power.

3.2 To what extent can we predict Knowledge Gain from Early Keystroke Patterns?

Following the statistical analysis, we evaluated the combined predictive power of the early keystroke features using machine learning models. For both the regression and classification tasks, a suite of different models was tested (see Tables 6 and 7). Although the primary goal of this study is exploratory, the top-performing model is conventionally selected during model selection. We therefore focus our reporting on the performance of the best performing model for each task in terms of the each evaluation metrics used. In the regression task (Table 6), the Random Forest Regressor exhibited the strongest performance among the evaluated models. It achieved the lowest root mean squared error (RMSE = 0.36). We normalized the RMSE using the observed range of the knowledge gain variable (−0.41 to 0.99). After normalization, the RMSE corresponds to 26% of this range. This means that, on average, the model's predictions are off by about one quarter of the total spread of knowledge gain values in the data. The same model also produced the highest correlation between predicted and observed knowledge gains ($r = 0.42$). For the classification task (Table 7), the Random Forest classifier was the top-performing model. It achieved an AUC-ROC of 0.68, indicating that—based only on keystroke patterns from the first two questions—the model has a 68% chance of correctly distinguishing between a randomly chosen eventual successful learner and a randomly chosen unsuccessful learner. The model also achieved an F1-score of 0.71, indicating a strong balance between precision and recall, and a balanced accuracy of 0.67.

Table 6: Regression results for predicting knowledge gain

Model	RMSE	r
Random Forest Regressor	**0.36**	**0.42**
Extra Trees Regressor	0.37	0.40
Decision Tree Regressor	0.39	0.29
Linear Regression	0.40	0.25
Ridge Regression	0.39	0.25
M5prime	0.38	0.37

Table 7: Classification results for predicting above/below median knowledge gain

Model	ROC	PR	F1	BACC	ACC
Logistic Regression	0.63	0.64	0.60	0.61	0.61
Decision Tree	0.66	0.63	0.71	0.66	0.67
Random Forest	**0.68**	**0.67**	**0.71**	**0.67**	**0.67**
RIPPER	0.49	0.53	0.26	0.51	0.49
K*	0.63	0.64	0.63	0.60	0.60

4. DISCUSSION

4.1 Association Between Early Keystroke Patterns and Knowledge Gain

Our results indicate that students who achieved higher knowledge gains engaged with the first two post-instruction mathematics problems in ways that differed systematically from their lower-gaining peers. Specifically, higher-gain students exhibited slower response construction (longer average seconds between keys, longer seconds per keystroke), greater variability in pacing (higher variation in seconds between keys), and more evidence of answer modification and revision (higher keystroke edit distance, correction density, and correction proportion). The observed effect sizes were small to moderate (Cliff's $\delta = 0.15$–0.38), with the strongest associations found for latency-based features. These patterns were consistent across multiple analytical approaches: group comparisons, correlation analyses, and quartile examinations all converged on the finding that slower, more variable, and more corrective early typing behaviors were associated with greater subsequent knowledge gain.

These findings suggest—though do not definitively establish—that students who achieved higher knowledge gains exhibited different behavioral patterns during their first two practice problems compared to lower-gaining students. One plausible interpretation is that slower, more variable typing with greater revision activity reflects more effortful cognitive processing. In mathematics problem solving, pauses between keystrokes could indicate time spent on mental calculation, deliberation about solution strategies, or verification of intermediate steps. Similarly, greater editing activity and a lower ratio of final answer length to total keystrokes may indicate that students reconsidered their initial responses rather than committing to their first attempt.

However, alternative interpretations must be considered. Slower, more variable keystroke patterns could also arise from factors unrelated to productive cognitive engagement, such as lower typing proficiency, confusion or uncertainty that may not always lead to learning, or individual differences in problem-solving tempo. Our analyses cannot distinguish whether these keystroke patterns reflect productive struggle that facilitates learning or whether both the keystroke patterns and subsequent learning are independently driven by underlying factors such as prior knowledge, metacognitive skills, or motivation. Students who already possess stronger foundational knowledge or better self-regulation may naturally exhibit more deliberate, revisionary behavior while also being better positioned to learn from practice. Without experimental manipulation, the observed associations remain correlational.

Moreover, while the correlations we observed were statistically significant, individual keystroke features explained only modest proportions of the variance in knowledge gain (3.6%–11.6%), suggesting that these behavioral signals capture some, but not all, of the factors that differentiate higher- and lower-gaining students. The moderate effect sizes similarly indicate that group differences, while consistent, are not absolute: there is substantial overlap in the keystroke distributions of successful and unsuccessful learners. This overlap suggests that keystroke patterns alone provide an incomplete picture of the learning process.

4.2 Predictive Value of Early Keystroke Patterns

The machine learning analyses provide additional perspective on the practical utility of early keystroke features for anticipating learning outcomes. In the regression task, the best-performing model (Random Forest) achieved an normalized RMSE corresponding to approximately 26% of the observed range of knowledge gain values and a correlation of $r = 0.42$ between predicted and actual gains. In the classification task, the Random Forest classifier distinguished between above- and below-median learners with an AUC-

ROC of 0.68, balanced accuracy of 0.67, and F1-score of 0.71. These results indicate that early keystroke patterns contain a detectable signal about subsequent knowledge gain, but the predictive accuracy is far from perfect.

To contextualize these findings, it is useful to consider what they mean in practical terms. An AUC-ROC of 0.68 indicates that, given two randomly selected students—one who will ultimately show higher knowledge gain and one who will show lower gain—the model has a 68% chance of correctly identifying which is which based solely on their first two problems' keystroke patterns. While this exceeds chance performance (50%), it also means the model will make incorrect orderings approximately one-third of the time. Similarly, the regression model's prediction error of 26% of the knowledge gain range suggests that individual predictions may deviate substantially from actual outcomes in many cases. The classification model's balanced accuracy of 0.67 further indicates that roughly two-thirds of students are classified correctly, leaving one-third misclassified. These levels of predictive performance are comparable to those reported in related early prediction tasks in educational contexts. Prior work predicting wheel-spinning from early problem attempts [4] and forecasting course outcomes from initial interaction data [6, 10] has similarly shown modest but above-chance accuracy. The present findings extend this pattern to the specific case of predicting knowledge gain during skill practice, confirming that early behavioral signals are informative but not deterministic of learning outcomes.

4.3 Implications for Adaptive Learning Systems

Despite the modest predictive accuracy, our findings may have practical value for adaptive instructional systems, particularly when considered in light of the costs and benefits of different types of errors. Prior work in educational data mining has largely focused on detecting sustained unproductive patterns, such as wheel-spinning, where students fail to progress toward mastery despite extended practice [4, 15]. These detection systems typically operate after students have completed many problems, by which point valuable learning time may have been lost. The present results suggest that behavioral differences between higher- and lower-gaining students may be detectable much earlier—specifically, within the first two problems following instruction, before extended practice has occurred. This temporal advantage creates potential opportunities for earlier adaptation, though several important caveats apply. First, the utility of early prediction depends on the availability of effective interventions. If an adaptive system identifies a student as being likely to make limited knowledge gains based on rapid keystroke patterns with minimal revision, the appropriate response is not obvious. Prompting such students to slow down, explain their reasoning, or verify their work might be beneficial if their direct response construction reflects superficial engagement, but could be counterproductive if it indicates genuine fluency or if the additional cognitive load interferes with their learning process. Conversely, students showing slower, more variable patterns with substantial revision might benefit from encouragement or targeted scaffolding, but only if their behavior reflects struggle rather than successful work that happens to be more deliberate.

Furthermore, the modest predictive accuracy means that system-level decisions based on these features should be designed to accommodate uncertainty. Rather than triggering interventions based on early keystroke patterns alone, adaptive systems might use these signals as one input among many—combined with correctness, prior performance, time on task, and other indicators—to inform more nuanced responses. For example, a system might use early keystroke patterns to adjust the probability threshold for offering optional support rather than relying on rigid decision rules. Alternatively, these patterns could inform the sequencing or selection of subsequent problems, steering students toward items that provide appropriate challenge based on estimated learning trajectories.

4.4 Contribution and Open Questions

From a theoretical perspective, our findings contribute to an emerging body of work suggesting that fine-grained behavioral traces, such as keystrokes, can provide insight into learning as it unfolds [1, 6]. The observed associations between early keystroke features and subsequent knowledge gain are consistent with the idea that behaviors reflecting deliberation, self-monitoring, and revision may be detectable in real time and may relate to learning outcomes. However, several important questions remain unresolved. First, the causal mechanisms underlying the observed associations are not established. It remains unclear whether the keystroke patterns themselves reflect cognitive processes that causally contribute to learning, or whether both keystroke behavior and learning outcomes are independently influenced by factors such as prior knowledge, motivation, or self-regulation. Second, the specific cognitive processes that give rise to different keystroke patterns remain inferential. While slower, more variable typing has been interpreted here as potentially reflecting planning or checking, alternative explanations such as retrieval difficulty or hesitation are also plausible. Finally, the temporal dynamics of keystroke behavior and learning are not well understood. Keystroke patterns may evolve across practice, and patterns later in learning may provide different or stronger signals than those observed at the outset. Longitudinal analyses could help clarify how keystroke behavior changes over time and how these changes relate to knowledge gain trajectories.

4.5 Limitations

Several limitations should be noted. First, the data were drawn from a single online mathematics platform, which may limit generalizability. Second, knowledge gain was operationalized using Bayesian Knowledge Tracing estimates, which are model-based inferences rather than direct measures of learning. While BKT is widely used and validated [11], measurement error may attenuate observed associations. Third, the analysis focused on a specific set of keystroke features. Other features capturing different aspects of response construction may provide additional information. Finally, this study examined association and prediction rather than intervention. Experimental studies are needed to determine whether instructional adaptations informed by early keystroke patterns can causally improve learning outcomes.

4.6 Conclusion

In summary, our findings show that early keystroke patterns during the first two problems of skill practice are as-

sociated with subsequent knowledge gain in mathematics, with higher-gaining students exhibiting slower, more variable, and more revision-oriented answer construction. Machine learning models trained on these early features achieved modest but above-chance accuracy in predicting learning outcomes. While these results are consistent with theoretical accounts emphasizing the distinction between performance and learning, they also highlight important limitations and uncertainties. Nevertheless, the findings suggest that fine-grained behavioral data from the earliest moments of independent practice may offer valuable information for understanding and supporting learning trajectories, motivating future work on causal mechanisms, intervention design, and generalizability across domains and contexts.

4.7 Acknowledgments

This work was supported by the Learning Engineering Virtual Institute (LEVI)

5. REFERENCES

[1] V. M. Baaijen, D. Galbraith, and K. de Glopper. Keystroke analysis: Reflections on procedures and measures. *Written Communication*, 29(3):246–277, 2012.

[2] R. S. D. Baker, A. T. Corbett, and V. Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. In *Proceedings of the International Conference on Intelligent Tutoring Systems*, pages 406–415, Berlin, Heidelberg, 2008. Springer.

[3] R. S. D. Baker, A. T. Corbett, S. M. Gowda, A. Z. Wagner, B. A. MacLaren, L. R. Kauffman, A. P. Mitchell, and S. Giguere. Contextual slip and prediction of student performance after use of an intelligent tutor. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*, pages 52–63, Berlin, Heidelberg, 2010. Springer.

[4] J. E. Beck and Y. Gong. Wheel-spinning: Students who fail to master a skill. In *International conference on artificial intelligence in education*, pages 431–440. Springer, 2013.

[5] A. F. Botelho, A. Varatharaj, T. Patikorn, D. Doherty, S. A. Adjei, and J. E. Beck. Developing early detectors of student attrition and wheel spinning using deep learning. *IEEE Transactions on Learning Technologies*, 12(2):158–170, 2019.

[6] K. Casey. Using keystroke analytics to improve pass–fail classifiers. *Journal of Learning Analytics*, 4(2):189–211, 2017.

[7] R. Conijn, C. Cook, M. van Zaanen, and L. Van Waes. Early prediction of writing quality using keystroke logging. *International Journal of Artificial Intelligence in Education*, 32(4):835–866, 2022.

[8] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1995.

[9] D. Galbraith and V. M. Baaijen. Aligning keystrokes with cognitive processes in writing. In *Observing Writing: Insights from Keystroke Logging and Handwriting*, pages 306–325. 2019.

[10] G. Gao, A. Leon, A. Jetten, J. Turner, H. Almoubayyed, S. Fancsali, and E. Brunskill. Predicting long-term student outcomes from short-term edtech log data. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 631–641, 2025.

[11] T. Gervet, K. R. Koedinger, J. Schneider, and T. Mitchell. When is deep learning the best approach to knowledge tracing? *Journal of Educational Data Mining*, 12(3):31–54, 2020.

[12] E. Kleinman, M. Shergadwala, Z. Teng, J. Villareale, A. Bryant, J. Zhu, and M. S. El-Nasr. Analyzing students' problem-solving sequences: A human-in-the-loop approach. *Journal of Learning Analytics*, 9(2):138–160, 2022.

[13] K. R. Koedinger, A. T. Corbett, and C. Perfetti. The knowledge–learning–instruction framework: Bridging the science–practice chasm to enhance robust student learning. *Cognitive Science*, 36(5):757–798, 2012.

[14] W. Ma, O. O. Adesope, J. C. Nesbit, and Q. Liu. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4):901–918, 2014.

[15] N. Matsuda, S. Chandrasekaran, and J. C. Stamper. How quickly can wheel spinning be detected? In *Edm*, pages 607–608. ERIC, 2016.

[16] L. Van Waes and M. Leijten. Fluency in writing: A multidimensional perspective on writing fluency applied to l1 and l2. *Computers and Composition*, 38:79–95, 2015.

[17] K. VanLehn. The behavior of tutoring systems. *International journal of artificial intelligence in education*, 16(3):227–265, 2006.

[18] C. Zhang, Y. Huang, J. Wang, D. Lu, W. Fang, J. Stamper, S. Fancsali, K. Holstein, and V. Aleven. Early detection of wheel spinning: Comparison across tutors, models, features, and operationalizations. *International Educational Data Mining Society*, 2019.