Temporally Rich Features Capture Variable Performance Associated with Elementary Students' Lower Math Self-concept

Shamya Karumbaiah[1], Jaclyn Ocumpaugh[1], Matthew J. Labrum[2], and Ryan S. Baker[1]
[1] University of Pennsylvania, Philadelphia, PA USA, [2] Imagine Learning, Provo, UT USA
shamya, ojaclyn@upenn.edu, matthew.labrum@imaginelearning.com, rybaker@upenn.edu

ABSTRACT: A better understanding of the relationship between self-concept in mathematics and fine-grained behavior logs from students' interactions with intelligent tutoring systems (ITSs) could help researchers better understand self-concept, which in turn could lead to improved designs in interventions intended to improve a student's self-concept. Yet, to date, learning analytics researchers have had only limited success in modeling these relationships. This exploratory study uses correlation mining to investigate the potential of temporally-rich features to capture variance in student performance. Results suggest detecting such inconsistencies in students' performance may be key to developing more robust models to infer self-concept, as well as to understanding how differences in it emerge among elementary students.

Keywords: self-efficacy, math identity, math performance, time-series, non-cognitive skill

1 INTRODUCTION

Self-concept has been shown to predict student achievement (Spinath et al., 2006) and appears to be conceptually related to other non-cognitive and motivational constructs including expectancy and value (Eccles and Wigfield, 1995), intrinsic interest (Gottfried, 1985) and intrinsic values (Pintrich et al., 1993). As summarized in Marsh et al. (2005), domain-specific self-concept (e.g., mathematics self-concept) shows developmental patterns of decline from early childhood to adolescence and then increases during early adulthood. The rich environment provided by many intelligent tutoring systems (ITSs), which can provide fine-grained assessment of behavior, affect, and cognition, might prove fruitful for better understanding the developmental changes in this construct. Indeed, Bernacki et al. (2015) found learners' self-efficacy (self-beliefs related to a specific task) varied reliably over an algebra unit in Cognitive Tutor in their investigation of the stability of self-efficacy and its relationship to problem-solving performance.

However, there has been limited research into how self-efficacy and self-concept relate to the types of behavior seen in intelligent tutors and other online learning systems. In part, this may be because self-concept survey instruments were deliberately devised to diverge from straight-forward measures of performance like grades or test scores (e.g., Gottfried, 1985). This may explain the limited success in correlating survey measures of self-concept to behaviors in ITSs. For example, Slater et al. (2018) used correlation mining to examine 185 features of student interactions (aggregated at a monthly and yearly level) with an ITS and found only 18 that showed a significant relationship with self-concept. That said, there were interesting patterns in these results, suggesting that future work should focus on engineering features that better captured variance in the students'

performance. Building on the promise of these findings, we present an exploratory paper where we investigate time-series features—which capture more complex and fine-grained characteristics of students' interactions with an ITS over long periods of time. As our data show, these temporally-rich features seem better able to capture the kind of variation needed to characterize math self-concept.

2 METHOD

2.1 Reasoning Mind Foundations

Reasoning Mind *Foundations* (Figure 1) is an intelligent tutoring system for math learning used by over 100,000 elementary school students in the United States including rural, urban, and suburban schools. Many of these students are from traditionally underrepresented populations. In this blended environment, students learn through self-paced problem solving, mathematical games, and interactive explanations. There are three main types of problems in this ITS based on the increasing levels of difficulty: 1) A-level problems on fundamental skills; 2) B-level (optional) problems on a combination of skills; and 3) C-level (optional) problems on higher order thinking skills.

Figure 1: *Foundations' home screen.*

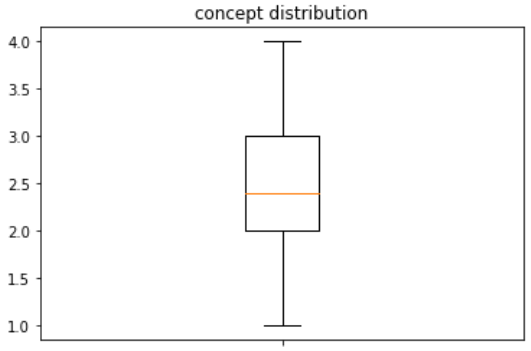

Figure 2: Distribution of average math self-concept scores of *Foundations* students.

2.2 Data

This study examines data collected in the 2017--18 school year from 2nd--5th grade students in Texas classrooms who interacted with *Foundations* as part of their regular mathematics instruction.

Math Self-concept Data - Surveys adapted from Marsh et al. (2005) (e.g, *Math just isn't my thing. Some topics in math are just so hard that I know from the start I'll never understand them.*) were administered at the end of the academic year 2017--2018 to collect 1566 students' self-reports on math self-concept using five items, each on a four-point Likert scale (Cronbach α = 0.74; mean = 2.42 (SD = 0.81); Figure 2).

Times-series Feature Extraction - Student performance on A-, B-, and C-level problems in *Foundations* was aggregated to engineer day-level sequences (time series) of the number of correct responses for each student (Figure 3). The average length of the times series is 70 days (SD = 35 days) spread out through the course of two semesters, with considerable differences in daily averages and standard deviations for the number of correct responses in different levels (A-level =

4.4 (4.79), B-level = 1.08 (3.42), C-level = 0.61 (2.51)). Time-series features were extracted with a Python package called *tsfresh* (Christ et al., 2018) which, in addition to providing high-level features describing meta-information of the time series, also calculates a comprehensive set of feature mappings that characterizes them. Across the three time series of the problem level performances, a total of 2382 features were extracted.

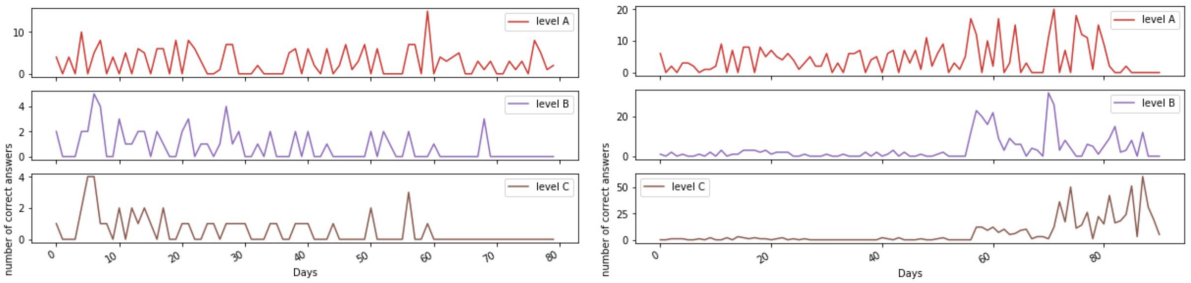

Figure 3: Example time series of the number of correct answers of level A, B, and C problems of a student with high post self-concept (*left*) and a student with low post self-concept (*right*). Note the difference in y-axis scales in the subplots.

3 RESULTS

A correlation analysis of the extracted time-series features with math self-concept yielded 113 significant correlations after using the Benjamini and Hochberg (1995) post-hoc procedure to control for false discovery (at $\alpha = 0.05$). All of the significant features were negatively associated with self-concept. Table 1 presents problem-type level aggregate correlations based on ten of the most frequent types of time-series features from the full correlation list[1]. As the descriptions in the table show, most of these features capture the greater variances in the correctness of math problems, a result comparable with Slater et al.'s (2018) findings that inconsistent performance is associated with lower math self-concept. For instance, the *change quantile* category, which captures the point-by-point change (mean and variance) in an ordered list of a student's daily math performance, contributes 31 of the 133 significant correlations found.

Other notable patterns emerge from these results. First, feature types (as described in Table 1) differ with respect to the problem levels that are most relevant (i.e., A-, B-, and C-level problems). All three levels emerge as significant for some feature categories, notably the *change quantile* category where A- and B-level problems each contribute 10 significant features and C-level problems contribute 11 more. In contrast, other feature categories are only significant for features derived from C-level problems; this includes *symmetry looking* and *CWT coefficients* (14 features each), as well as *approximate entropy* (5 features), *FFT coefficients* (3 features), and *Ratio beyond R Sigma* (3 features)---a greater description of these features is given in Table 1. Overall, the majority of the significant correlations (78/113) involved features extracted from C-level problem performance, which may be related to the fact that students have more autonomy in their decision to complete C-level problems (i.e., they are usually optional). There are also differences, within certain feature categories, in how those features are operationalized. For example, *change quantile* features

[1] A full list of all significant correlations after controlling for false discovery is available at [link redacted for review]

involving A- and B-level problems rely on earlier quantile ranges of the time series (i.e., 0.1--0.6 and 0.2--0.8), while those involving C-level problems make greater use of later quantiles of the time series (i.e., 0.0--1.0 and 0.4--0.8). That is, for the easier problem sets (A- and B-level problems), self-concept is negatively associated with variance that occurs in the lower ranges of the number of correct responses, while with the more challenging (and optional) C-level problems, the variance is more likely to be relevant in the higher ranges of the number of correct responses.

Table 1: Features categories of the time series (TS) with significant correlation with self-concept.

Category	Counts			mean R	Description[2]
	A	B	C		
Change quantiles	10	10	11	-0.16	Average, absolute value of consecutive changes inside different quantile ranges of TS. Higher change corresponds to a higher inconsistency in student performance.
Symmetry looking	0	0	14	-0.14	Boolean value specifying if the distribution symmetric: Is |mean(TS)-median(TS)| < r * (max(TS)-min(TS))? If symmetric, the values in TS occur regularly.
CWT coefficients	0	0	14	-0.15	Coefficients of the continuous wavelet transform for the Ricker wavelet; these give information about the amplitude of TS and how that amplitude varies over time.
Quantile	3	3	4	-0.16	*q*th quantile value of TS.
Descriptive statistics	1	1	5	-0.15	Mean, median, maximum, mean absolute change, and if SD(TS) > r * (max(TS)-min(TS))?
Number of peaks	0	2	5	-0.17	Number of peaks in TS subsequences, where peaks signify a sudden increase in the number of correct responses.
Approximate entropy	0	0	5	-0.18	Amount of irregularity and unpredictability of fluctuations in different TS ranges. Higher entropy corresponds to lower regularity in student performance.
Sum of recurring data points	1	1	1	-0.15	Sum of values in TS that are present more than once (e.g., if a student had 12 correct responses on two or more different days, 12 would be counted towards this feature).
FFT coefficients	0	0	3	-0.15	Fourier coefficients of the one-dimensional discrete Fourier Transform; these give information about the underlying periods in the TS (e.g., weekly or monthly periods of performance).
Ratio beyond R sigma	0	0	3	-0.15	Ratio of student performance values that are farther than R*SD(TS) away from the mean(TS), so that larger values show greater variance.

4 DISCUSSION

Learning analytics research has had limited success in modelling asynchronous survey measures of non-cognitive constructs (e.g., self-concept) from student behaviors in an ITS, compared to other

[2] The detailed feature description is at https://tsfresh.readthedocs.io/en/v0.11.1/text/list_of_features.html

constructs. In contrast to previous research in this area, our study explores time-series features which capture finer-grained and longer-term temporal variations in student performance than often captured in the feature engineering process. We have demonstrated that these time-series features look promising as indicators of elementary students' math self-concept. The primary takeaway of our analysis is the negative relationship between inconsistencies in student performance and their domain-specific self-concept. Thus, an immediate implication of this work is to further examine the rate at which the different problems are being introduced to students in the ITS. C-level problems, which are more difficult and optional, show the most promise for predicting self-concept. Particularly, there seems to be a delay in attempting harder (C-level) math problems (see Figure 3) among students with lower self-concept. This calls into attention the intentionality of the students' behaviors, suggesting that one way to test for self-concept is to give students optional opportunities for more challenging practice. Our future work will use these findings to better design new features, to develop a predictive model of math self-concept for students using *Foundations,* and eventually to design interventions for students with low self-concept.

REFERENCES

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37(2):122.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bernacki, M. L., Nokes-Malach, T. J., and Aleven, V. (2015). Examining self-efficacy during learning: variability and relations to behavior, performance, and learning. *Metacognition and Learning*, 10(1):99–117.

Christ, M., Braun, N., Neuffer, J., and Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 307:72–77.

Eccles, J. S. and Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin*, 21(3):215–225.

Gottfried, A. E. (1985). Academic intrinsic motivation in elementary and junior high school students. *Journal of Educational Psychology*, 77(6):631.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., and Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76(2):397–416.

Pintrich, P. R., Smith, D. A., Garcia, T., and McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement*, 53(3):801–813.

Slater, S., Ocumpaugh, J., Baker, R., Li, J., and Labrum, M. (2018). Identifying changes in math identity through adaptive learning systems use. In *Proceedings of the 26th International Conference on Computers in Education*.

Spinath, B., Spinath, F. M., Harlaar, N., and Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. *Intelligence*, 34(4):363–374.