EasyChair Preprint

№ 768

Empowering Tutors with Big-data Learning Analytics

Uma Vijh, Josine Verhagen, Webb Phillips and Ji An

February 5, 2019

Empowering Tutors with Big-data Learning Analytics

Author(s): Uma P. Vijh
Kidaptive Inc.
uma.vijh@kidaptive.com

Author(s): Josine Verhagen
Kidaptive Inc.
josine.verhagen@kidaptive.com

Author(s): Webb Phillips
Converz Analytics B.V.
webb.phillips@gmail.com

Author(s): Ji An
Kidaptive Inc.
ji.an@kidaptive.com

ABSTRACT: **Presentation. Online education has been growing over the past few years, and massive amounts of learning data are being generated. We are reporting on our efforts to use learning analytics to empower teachers to help all learners reach their full potential. We provide teachers with insights about student behavior and achievement on a weekly basis and supplement these with summary monthly reports about student study patterns and trends. This paper provides more detailed descriptions of these reports and also includes preliminary efficacy study results that show positive effects on mean student test scores.**

Keywords: study pattern analytics, teacher insights, informal learning environments, predictive modeling, bayesian item response theory

1. INTRODUCTION

In recent years, learning analytics has emerged as a powerful learning tool for teachers who participate in online learning programs. Big-data learning analytics deciphers massive amounts of data generated in different learning contexts. It can help to assess students' academic progress, predict their future performance, and identify potential problems (Johnson, Adams & Cummins, 2012). For teachers, learning analytics can be used to carry out a more in-depth analysis of the teaching process to provide more targeted teaching interventions for students (Chen, Heritage & Lee, 2005).

2. BACKGROUND

In this paper, we describe our learning analytics efforts to support teachers helping K–6 learners. The data are event data as learners interact with curricular content from a Korean partner's tablet-based educational system. The system supports over 200,000 learners in Math, Korean, Social Studies and Science, following the Korean national curriculum. Students in the program mostly work at home and are visited by a teacher once a week. The

content is arranged in weekly topics and further broken down into small content blocks containing lectures and practice questions. Each week ends with a test. As the learners progress through the curriculum, they watch lectures, answer 50–100 practice questions, and complete a test with 10–20 questions. Our technology provides teachers with weekly reports that are updated continuously, as well as monthly reports to track the learners' progress over time. These reports (described in subsequent sections) contain more information than just the correctness/incorrectness of student answers. Our cloud-based analytics engine processes millions of events streaming in, using regularly calibrated psychometric models to produce hundreds of distinct personalized metrics and insights. These insights are dynamically prioritized, with the most important passed along to teachers to help all learners reach their full potential.

3. METHODS

3.1. Description of the report:

For every weekly curricular unit attempted by a learner, we produce a report for that learner's teacher. In this weekly report, we provide general behavioral insights, specific question-level insights, and one overall message about the learner's behavior and achievement during the week.

The behaviors analyzed are: skipping questions, answering too quickly or slowly, guessing, leaving parts of the question blank, skipping a question after getting the previous one wrong, retrying or not retrying incorrect questions, watching or not watching all lectures, and checking or not checking hints after getting a question wrong. In addition to these behavior metrics, the reports also include question insights based on personalized speed and ability estimates and performance on the weekly test. These details empower the teacher to quickly identify questions/concepts each student is struggling with, praise good study habits, and assess student performance not only at a personal level but also in comparison with peers.

To tell teachers more than whether question responses were correct, we developed some additional insights about responses.

3.1.1. Answer speed:

An item is flagged as answered relatively fast or slow based on the learner's expected time on the item given their working speed and whether the learner is answering faster or slower than 90% of the other students answering the item. Based the learner's history in a given subject, a Bayesian personalized estimate is kept of his or her working speed. The working speed is updated only based on items the student answered correctly, to keep the estimate from plummeting when a student is just skipping through questions. The estimate is based on a linear mixed model of the logarithm of the response time, with the learner's working speed estimate calculated relative to the average time spent on the item by other learners. E.g., if a learner's response time is faster than 90% of other learners' response times but consistent with that learner's working speed, the item is not flagged as too fast.

3.1.2. Item difficulty:

Based on the learner's ability estimate and question difficulty, questions are categorized as hard (<50% probability of getting the question correct), easy (>80% probability of getting the question correct) or medium for a given learner. Ability estimates are based on an adjusted version of Bayesian Item Response Theory models (Bock & Mislevy, 1982; Van der Linden & Glas, 2000) developed for adaptive testing, which allows the ability estimate to be updated after each question. Because reports are generated on an edition level, the final ability estimate and question difficulty estimates represent how well a learner did compared to other learners at the end of that edition. At the start of each edition, the prior probability distribution is set to the average of the priors from the three previous editions, with a wide standard deviation to allow for a different ability level for the topic at hand.

3.1.3. Guessing:

We developed a general model for estimating thresholds for response times that are short enough to suggest that students probably guessed the answer (See Wise & Kong, 2000; Baker et al. 2006 for discussion on rapid response times). This model applies across all question types and is based on the distribution of response data and corresponding pass-rates on a per-question basis. Using this model, we were able to categorize responses as "guessed" much more accurately than simply setting an arbitrary response time for all questions. Comparing response times to pass-rates, most questions have a region of low response times with low pass-rates and a region of higher response times with higher pass-rates. Then the pass-rate gradually declines for even higher response times. The log normal distribution shares a similar shape, and therefore makes a good function to model response time vs. outcome. As an example, Figure 1 shows a model for one math question after having optimized four coefficients. These models have low mean squared error (~0.05) compared to actual response time vs. outcome data. We found that our model needed at least 50 correct and 50 incorrect responses to be reliable. Of the 58,806 questions for which our analytic platform had responses and response times, our modeling algorithm assigned a

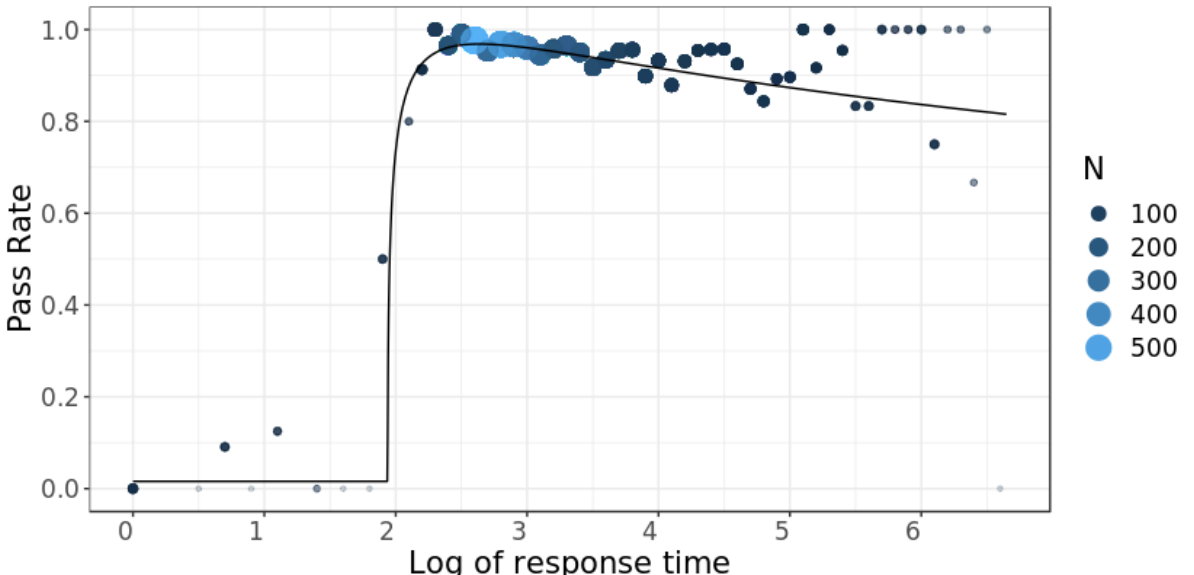

Figure 1: Guessing model using response times for one question in the math curriculum

default guessing threshold of up to one second for 69% of questions, specific thresholds greater than one second for 29%, and no guessing threshold for 1.6% (these were cases in which the percentage correct at one second was almost as high as or higher than the percentage correct at the middle 20% of response times for correct answers).

The combination of personalized answer speed, item difficulty, and item correctness produces insights regarding sets of items. Based on the individual ability estimate and the estimated item difficulty of the items in the next test in the curriculum for each learner, we also record an estimate of that learner's predicted performance on the upcoming test. We then use this estimate to provide further insight to the tutors (e.g., to congratulate or encourage the learner to do their best).

As learners work through the curriculum, we also provide monthly reports to the tutors, summarizing the learner's activity for the month as well as trends across months. This helps the tutor evaluate student learning and growth, praise improving study behaviors, and celebrate achievements.

4. EVALUATION OF INTERVENTION

Our reports were provided to all users of our partner's platform, so a direct control group was not available for the evaluation of the program. We evaluated efficacy of the product in two ways using linear mixed models:

1. *A Difference in difference analysis of historical data*: We compared the differences in test scores in the current year with those in the previous year, before and after launch of the teacher reports. This method accounts for the seasonal differences in course material, but the individual students are of course different. To mitigate this we included the random effects for difficulty of particular curricular material and individual learner ability. Month and year were modeled as fixed effects, and we also included interaction

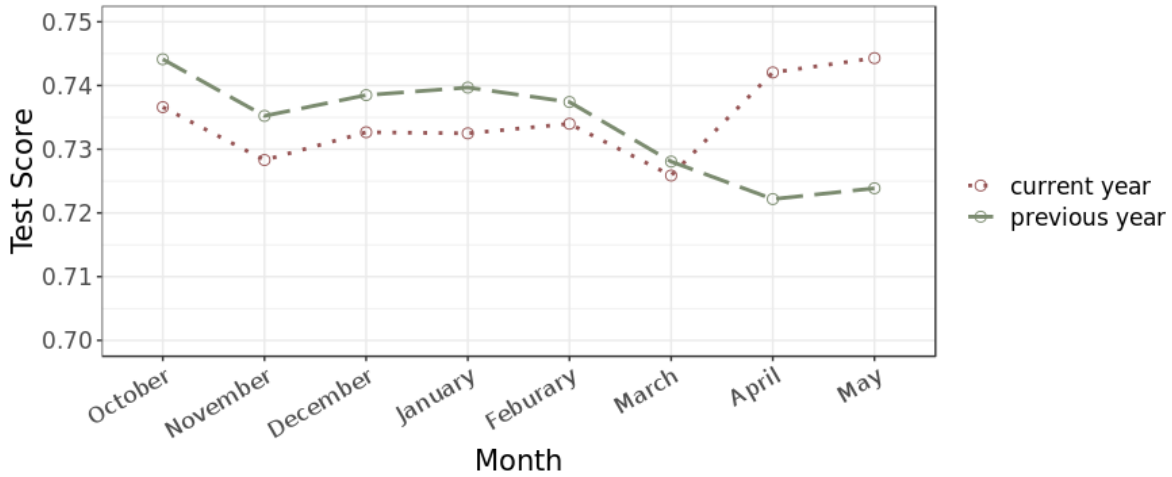

Figure 2: Mean test scores for Korean, for the current year and the previous year. Our intervention program was implemented in February.

effects between month and year. We used over 1.2 million individual scores of ~40,000 learners for the subject Korean. We included data for eight months from each year; the

data selected for this year covered three months before the implementation of our tutoring support service (February 2018) and five months after. Figure 2, shows the mean test scores for Korean during the previous year (2017) and the current year. The differences between current year and the previous year were essentially constant until January, with overall test score in the previous year being slightly higher than the current year. Starting from February, the current year scores start to catch up, and by April they outperform the previous year's scores. The increase ranged between 0.4 and 3.6 points across all the subjects, Math, Korean, Social Studies and Science on the scale of 0–100 points. Statistically significant, positive interaction effects start around one month after the implementation of the program, indicating that the test scores relative to last year have shown improvement after the start of the service.

2. A analysis based on frequency of report utilization by teachers: The historical analysis

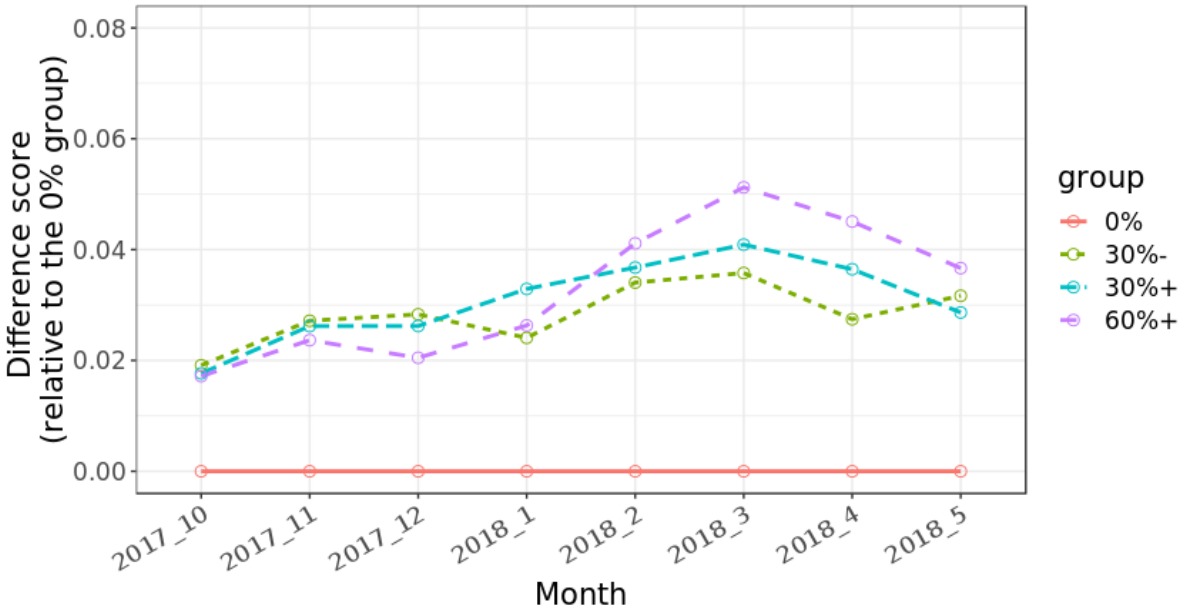

Figure 3: Differences in the mean test scores of students whose reports were viewed, grouped by the fraction of their reports viewed.

does not measure a direct effect of intervention by the teachers who are empowered by our reports. To evaluate a more direct effect we compared the test scores of students as a function of the rate at which their reports were viewed by the teacher. Our hypothesis was that teachers empowered by the personalized insights would provide timely intervention and over time positively influence student behavior. In addition to the year/month fixed effects and learner/material random effects, we grouped teachers based on what proportion of the students' reports they viewed and used the resulting group membership as a fixed effect to model the test scores. We analyzed all the subjects, Math, Korean, Social Studies and Science for which we provided reports. Figure 4 shows the difference in scores of the students whose teachers viewed their reports at different frequencies compared to those whose reports were not used at all, for Math. The performance of students whose teachers are in the never-viewed group is clearly worse than that of students whose teachers are in the three sometimes-viewed groups,

but those differences increase after January 2018, indicating an improvement associated with teachers viewing reports. As shown in Table 1, the improvements in scores relative to the never-viewed group range from 1.02 to 3.07 points depending on the rate of teacher viewing and time of the year. The statistical significance of these differences is indicated in parentheses and explained in the footnote.

Table 1: Differences in test scores between the indicated group and the group without any report views. Statistical significance indicated in parenthesis[1]

% of reports viewed	2018/1	2018/2	2018/3	2018/4	2018/5
< 30%	0.72(*)				
30% - 60%	0.52(.)	0.65(*)	0.73(*)	0.81(**)	0.93(**)
> 60%	0.98(***)	1.12(***)	1.20(***)	1.69(***)	1.40(***)

5. CONCLUSION

In this paper we have described a real-world instance of learning analytics indirectly supporting ~200,000 learners through personalized weekly and monthly reports sent to those learners' teachers. These reports characterize a variety of learning-relevant study behaviors to help teachers identify and correct bad habits, praise and reinforce good habits, and optimally direct each learner's study efforts. Two types of analyses comparing scores before and after implementation of personalized insights to tutors suggest a positive effect on test scores, especially for students whose reports are frequently viewed by their teachers.

REFERENCES

Baker, R., Koedinger, K. R., Corbett, A. T., Wagner, A. Z., Evenson, S. et al.. Adapting to When Students Game an Intelligent Tutoring System. International Conference on Intelligent Tutoring Systems, 2006, Jhongli, Taiwan. 2006. <hal-00190177>

Becker, S.A., Cummins, M., Davis, A., Freeman, A., Glesinger Hall, C. & Ananthanarayanan, V. (2017). NMC Horizon Report: 2017 Higher Education Edition. Austin, Texas: The New Media Consortium.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied psychological measurement, 6(4), 431-444.

Chen, E., Heritage, M. & Lee, J. (2005). Identifying and Monitoring Students' Learning Needs With Technology. Journal of Education for Students Placed at Risk, 10 (3), 309-332.

Wise, S. L. & Kong, X. Applied Measurement in Education, 2005, Montreal, April, 2005 Van der Linden, W. J., & Glas, C. A. (Eds.). (2000). Computerized adaptive testing: Theory and practice. Dordrecht: Kluwer Academic.

[1] significance: . p < .1; * p< .05; ** p < .01; *** p < .001