Educational Equity Through Combined Human-AI Personalization: A Propensity Matching Evaluation

Danielle R. Chine[1[0000-0001-8196-3252]], Cassandra Brentley[2], Carmen Thomas-Browne[2], J. Elizabeth Richey[2[0000-0002-0045-6855]], Abdulmenaf Gul[1], Paulo F. Carvalho[1[0000-0002-0449-3733]], Lee Branstetter[1[0000-0001-7835-0527]], Kenneth R. Koedinger[1]

[1] Carnegie Mellon University, Pittsburgh PA 15213, USA
{dchine, pcarvalh, branstet, kk1u}@andrew.cmu.edu menafgul@gmail.com
[2] University of Pittsburgh, Pittsburgh PA 15260, USA
{cassandrabrentley, cgt9}@pitt.edu
jelizabethrichey@gmail.com

Abstract. Recent developments in combined human-computer tutoring systems show promise in narrowing math achievement gaps among marginalized stu dents. We present an evaluation of the use of the Personalized Learning^2, a hy brid tutoring approach whereby human mentoring *and* AI tutoring are combined to personalize learning with respect to students' motivational and cognitive needs. The approach assumes achievement gaps emerge from differences in learning opportunities and seeks to increase such opportunities for marginalized students through after-school programs, such as the Ready to Learn program. This program engaged diverse middle school students from three schools in an urban district. We compared achievement growth of 70 treatment students in this program with a control group of 380 students from the same district select ed by propensity matching to have similar demographics and prior achievement. Based on standardized math assessments (NWEA Measures of Academic Pro gress) given one year apart, we found the gain of treatment students (6.8 points) was nearly double the gain of the control group (3.6 points). Further supporting the inference that greater learning was caused by the math-focused treatment and not by some selection bias, we found no significant differences in reading achievement between treatment and control participants. These results show promise that greater educational equity can be achieved at reasonable costs through after-school programs that combine the use of low-cost paraprofession al mentors and computer-based tutoring.

Keywords: Personalized learning, Cognitive tutoring, Design-based research

1 Introduction

The impact of combined human mentoring and AI-driven computer-based tutoring on student performance is encouraging, with an expanding stream of research showing promise in improving learning gains, especially in mathematics [8, 24]. AIED technologies involving human-computer teaming can lower the financial cost of personal ized tutoring and increase student learning [3, 19, 24]. The human mentors in these teams generally require less professional and on-the-job training than classroom teachers [8], which keeps human resource costs low. However, these mentors need additional support in providing personalized resources and skills development to as sist with specific student's needs. The use of human-computer teaming, particularly in the wake of the COVID-19 pandemic, gives mentors access to individualized re sources using AI-software based on students' existing math learning software and mentor input and feedback. We present results (i.e., EdTech usage, math and reading learning gains and outcomes) from the deployment of an after-school learning support system that integrates human mentors and AI tutoring (e.g., [17, 23]), with the aim of substantially reducing income and racial gaps in learning opportunities and outcomes. The Personalized Learning (PL^2) approach intends to maximize the synergies between the motivational capability of human mentors and the ability of computer-aided learn ing systems to provide low-cost personalized learning in pursuit of more equitable educational outcomes.

1.1 Related Work

Narrowing the Opportunity Gap. Marginalized students lack the means to access quality instructional services and experience lesser opportunities for learning [24] creating an opportunity gap. We define marginalized learners as, "students systemati cally denied equitable access to the same opportunities theoretically available to all students (p. 216)" often due to socioeconomic status, disability, or racial minoritiza tion among other factors [12]. Racial and economic learning gaps are preventing mil lions of American students from realizing their potential which perpetuates inequali ties of income and opportunity across generations [2]. Recently, the COVID-19 pan demic has exacerbated these inequalities with lower student achievement at the start of the 2021-22 school year (9 to 11 percentile points on standardized achievement assessments) than previous years hitting marginalized groups the hardest— minority students experiencing high-poverty [15]. Although achievement was lower across all groups, the achievement gap is present now more than ever with higher achieving and non-marginalized students making gains consistent with projected normative growth and marginalized, often under-achieving, students lagging behind further exacerbating the learning gap [15]. While these are recent and long-standing problems, researchers have struggled to identify effective solutions. Recent research undertaken in the Chi cago Public Schools in some of the city's highest-poverty neighborhoods, provides new grounds for hope [3, 7, 11]. Using a randomized control trial consisting of 2,700 students of whom 95% were Black or Hispanic, they demonstrate that just one year of intensive, personalized tutoring can narrow racial achievement gaps in mathematics by as much as one third. These gains come at a substantial cost. With one tutor providing instruction to just two students per class period, the cost exceeds the threshold of political feasibility in many districts, despite its proven efficacy.

Offering Low-Cost Tutoring. Advances in computer-aided learning provide a method of substantially lowering the cost of personalized tutoring, while maintaining the magnitude of the learning gains. Research on AI-driven computer-based tutoring has shown computer tutors can substantially accelerate student learning, especially in mathematics. In one recent large-scale randomized control trial, this technology was shown to double the rate of math learning [21]. Setren [25] showed that the use of another commercially available tool (eSparks), has a positive effect on learning gains for all students and can contribute to reducing inequality. Similarly, Muralidharan et al. [19] showed that a personalized technology-aided after-school program was successful in generating large learning gains among under-achieving students in a developing country. Despite positive findings, many students do not partake in the practice opportunities provided. We propose human mentoring to help motivate students to participate and to round out their learning experiences.

Supporting the Whole Child. While computer tutors can often provide effective support for student thinking and learning, these systems do not provide human sup port for social motivational development such as self-efficacy building [26], feelings of belonging [31], growth mindset [32], and valuing utility of STEM [10]. Using the last as an example, motivational support for students and parents to better appreciate the value of STEM learning produced about 50% greater achievement and future course enrollment, especially for low-performing and underserved students. Our pro posed intervention supports the whole child similar to Guryan et al. [7] in attending to the social-motivational needs and relationship building which is particularly important in middle school years. Milner [18] posits that to foster excellence, a learning environment should center on building and cultivating relationships with students. The synergy of human and computers has been studied in a similar fashion with the use of trialogues (the interaction of two agents with a human student) to address pedagogical goals and student's emotional state [6] and the use of tutorial dialogue agents to increase learning gains [14]. Similarly, peer-to-peer interaction within intelligent tutoring systems to scaffold learning has been researched via adaptive collaborative learning supports for both improving content learning and collaboration [30].

2 The Personalized Learning^2 Approach

Introduced in Schaldenbrand et al. [24], PL^{2}is a learning app that syncs with stu dents' existing math learning software and mentor input and feedback to improve students' math achievement.[1] This paper presents an evaluation of the general PL^2 approach, as both a learning app and tutoring method, to human-computer teaming for motivational mentoring and content tutoring. By combining research-driven mentor training with AI-powered software, the PL^{2}approach improves mentoring efficiency by connecting mentors to personalized resources with the click of a button. This con nection is achieved by a web app used by mentors and mentor supervisors. The PL^2 approach serves out-of-school tutoring programs, which choose a computer-based math tutoring system for students to use. The data from student interactions is passed to the PL^2 web app to power mentor decisions. Mentors make post-session reflections based on reports of student effort and progress. Mentors work with students to set or modify intermediate effort goals, much like the 10,000 step goals in physical fitness apps, such as doing 40 minutes of math practice a week. When students are missing effort or progress goals, the PL^{2}app provides suggestions for resources that the mentor can use themselves or with students to enhance student motivation, cognition, or metacognition.

PL^2 has been integrated with several math EdTech systems [24]. The two used in the evaluation we report on were MATHia and ALEKS. MATHia (formerly Cogni tive Tutor) uses a cognitive model of student problem solving to implement the model tracing and knowledge tracing algorithms for personalized tutoring [23]. It has been demonstrated to improve student learning in large-scale randomized field trials (e.g., [21]). ALEKS is an intelligent tutoring system based on knowledge space theory and it too has been demonstrated to improve student learning [17].

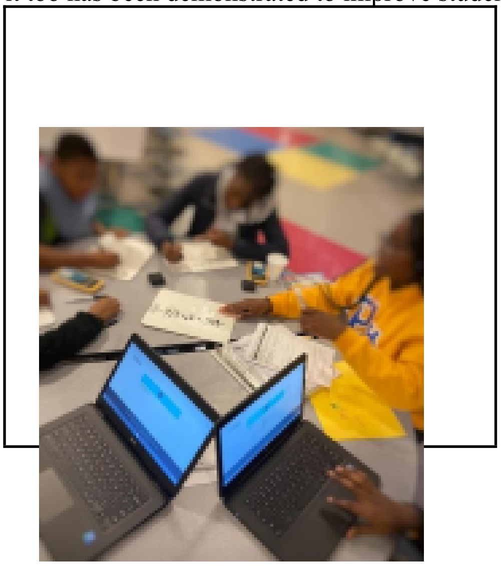


Fig. 1. The Ready to Learn program 2-hr in-person format commences with fellowship, followed by rotations between AI and in-person tutoring (40 min. each), and ends with reflection and team building. Students shown working with an in-person mentor and AI tutors.

2.1 Description of the Program

The Ready to Learn (RtL) program is offered by the Center for Urban Education at the

[1] http://personalizedlearning2.org/

University of Pittsburgh (Pitt). The program is a combination tutoring-mentoring initiative that connects students from Pitt and Carnegie Mellon University (CMU) with middle school students in select urban public schools to provide mentoring and math tutoring in an out-of-school context. The overarching goal of the RtL program is to provide students, especially students from disadvantaged contexts (i.e., living in poverty, experiencing racial bias, being part of a marginalized group) with experiences to support their academic improvement in mathematics. Since 2019, CMU re searchers have partnered with Pitt's Center for Urban Education to provide personal ized math mentoring. Small scale pilot versions of RtL in the spring and summer of 2019 paved the way for full implementation in the 2019-2020 academic year and following summer. RtL combines small group lessons, individual mentoring with trained CMU and Pitt undergraduate students, student engagement with adaptive AI driven learning software, and use of the Personalized Learning^{2}app to help mentors work together with technology to customize the learning experience for each student. The in-person session format (see Fig. 1) consisted of both a human personalized math lesson and personalized computer-based tutoring provided by the MATHia in telligent tutoring system. Because of the COVID pandemic, the program moved online in the summer of 2020. At the same time for logistical reasons, the computer based tutoring was changed to ALEKS. RtL builds student math confidence and com petence at no cost to students or to partnering schools. In the evaluation we describe below we included students that participated in an RtL program between the available assessments, that is, during the spring and summer of 2020.

By relying on undergraduate mentors and off-the-shelf math learning software and keeping the price of subscriptions to the PL^{2}app at reasonable levels, future imple mentations may be able to deliver learning gains for a modest cost per additional stu dent. Our calculations suggest a mentor cost of $360 per student/year.[2] With the addi tion of an annual EdTech license cost per student (i.e., $27 Mathia, $179.95 for ALEKS) and an annual PL^{2}student fee ($10), the cost of our intervention becomes $397-$550 per student. Thus, a marginal cost of about $500 per student is attaina ble—a fraction of the $3,500-$4,300 per student for other high-dosage tutoring pro grams [8].

3 Method

Participants in treatment and control groups included students mostly entering grades 6-7 at the 2019-2020 school year from three schools located in a medium-sized, ur ban, Pennsylvania school district. Two of the three schools have a higher proportion of disadvantaged students compared to the district aligning with the goal of RtL of reaching marginalized groups. The majority of students were in 6th (57%) and 7th grade (33%) grade. Students' demographics are summarized in Table 1. Among the treatment, approximately 74% were Black with an approximately equal gender distri bution (48% female). Most of the participants were eligible for free or reduced lunch (71%) and 20% were receiving an Individualized Education Program (IEP), which is a special education service.

Students' achievement was measured by the NWEA Measures of Academic Progress (MAP) assessment which the district administered a few times per year assessing students' math and reading achievement. In our evaluation, MAP scores for fall and winter of 2019-20 were used as pretest scores and MAP scores for fall and winter of 2020-21 were used as posttest scores. We used all four test scores to maximize the number of students for which at least one pre and one post score was available (see missing data discussion below). We note concerns that majority-based norming of standardized tests can create cultural bias and may exaggerate achievement gaps [13]. At the same time, we note efforts to reduce bias in standardized testing in gen eral [22] and that the questions on this test are representative of important learning goals for students (e.g., using rational numbers to solve real-world problems).

Table 1. Demographic group percent distribution (and number) demonstrates about 3/4 of participants belong to marginalized groups (i.e., Black and low SES)

Group Demographic Treatment (n) Control Gender Female 48% (34) 52%
(199) Race Black 74% (52) 81% (306) White 20% (14) 11% (43)
Hispanic 3% (2) 5% (19)
Free/Reduced Lunch Yes 71% (50) 79% (301) IEP Status IEP 20% (14) 17%
(64)

Control Group Creation via Propensity Matching. Toward our goal to evaluate whether extra learning opportunities provided by the PL^{2}approach enhanced student learning, we created a matched control group of similar students who did not receive these opportunities. The district provided anonymized score and demographic data from a total of 20,628 students across all grades for academic years 2019-20 and 2020-21. This data provided scores and demographics for the 72 students that participated in the PL^{2}treatment. These treatment demographics and pre-test scores were used as input into a rigorous propensity matching process to select a set of students as a control who were as similar as possible in demographics, grade level, and pre-test scores. An optimal full matching method [9, 27] was used to match each treatment student with multiple matching control students. Initially, all demographic factors were used to match students. However, gender and socioeconomic status (determined by free and reduced lunch designation) were found to be non-significant factors to balance groups and were removed from matching criteria. In the final matching, 70 students out of 72 in the treatment group were matched with 380 control students based on grade level, race, IEP (Individualized Education Program) status, and pretest math scores. The two students dropped from the treatment had a combination of these features for which there was no adequate match. Grade level was defined as an exact matching factor and a clipper value was defined for pretest math scores to ensure that matched units are close enough in terms of pretest math scores. In addition to propen

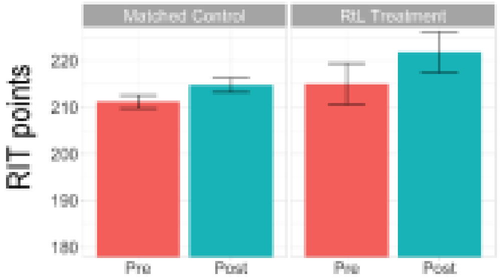
[2] Estimated mentor cost: $15/hr.*2 hrs./session*2 sessions/week*24 weeks/4 students per mentor =$360 per student/year.

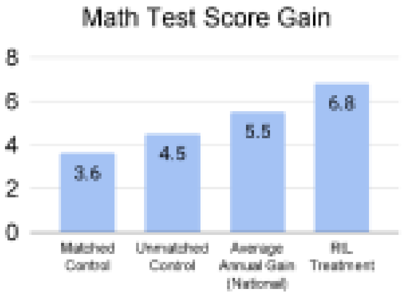
sity score matching, manual matching using exact matching of grade level, gender, race, socioeconomic status, IEP group, and math score within a 3-point range was used. Manual matching replicated all significant differences reported below with simi lar magnitudes.

Pandemic, Missing Data, and Imputation. Especially because of the pandemic, some participants completed only one of the two pre or posttests, mostly in fall 2020. We started with a total of 13,554 students in the district with at least one pretest and post test test score. The test with the highest missing rate was fall 2020 (14.9% control, 15.3% treatment). Missing data for all other tests ranged between 0% and 8.3%. Miss ing data were imputed using a single deterministic imputation model based on linear regression using the MICE package in R [29]. In this method, all demographic factors were entered as predictors and plausible synthetic values were generated for incom plete test scores. Incomplete math and reading scores were imputed separately. In addition to the single imputation method, multiple imputation based on stochastic regression with random error added to predicted values was tested. This method was tested with 20 samples. The results of both single and multiple imputation were iden tical with respect to all statistical threshold judgements. Since multiple imputation produces different matched samples, we present subsequent results using the single matched sample resulting from the single imputation method.

4 Results

Math Learning Gains. Figure 2a summarizes pre-to-post learning gains for the treatment group (rightmost bar) and a matched demographic control (leftmost bar) as well as two other reference points, a national average gain and a non-matched grade level control. On average, students in the treatment group grew 6.8 Rasch Interval Unit (RIT) points from pre to posttest, compared to 3.6 points for students in the matched control. NWEA MAP reports a typical one-year average growth as 5.5 RIT points [20].





(a) (b)

Fig. 2. a) Math test score gain comparison. b) Mean pre and posttest results. Both (a) and (b) illustrate the substantial gain for RtL participants compared to the matched control.

Further contextualizing this difference, we calculated the average growth in the same period for a non-matched control group, which included all students in the same grade without accounting for demographics. The non-matched control group showed a growth of 4.5 points in the same period, higher than the matched control, suggesting that the RtL program met its goal of working with disadvantaged students. In addi tion, the results may indicate evidence of the "COVID slide" among the control group given that the 3.6-point gain and grade level, "non-demographic" group gain of 4.5 points, are significantly lower than the average MAP expected annual growth. This aligns with the pandemic-related lag in mean growth evidenced by Lewis et al. [16] with RIT scores ranging approximately 2-3 points lower than pre-pandemic years dependent upon grade level and test administration. Past performance data of the par ticipating schools in comparison with national growth gains are needed to confirm evidence of "COVID slide." The most striking finding was the substantial growth differences among RtL participants with students performing significantly over the normative one year's growth.

Figure 2b provides average pre and posttest results for the matched control and RtL treatment. To assess whether the RtL treatment had a statistically greater pre-posttest gain than the matched control, we performed a repeated measures regression analysis. The model predicted MAP scores using group (experimental vs. control), test-time (pre vs. post) and the interaction between the two variables. As can be seen in Table 2, all students performed better in the posttest tests than the pretest tests, $\beta = 3.60$, t (448) = 8.30, $p < .0001$. Importantly, the interaction between type of test and group was also statistically significant, $\beta = 3.21$, t (448) = 2.92, $p = .0004$. MAP scores for students in the treatment group improved considerably more than for students in the control group (effect size of $d = 0.40$). An analysis of density plots of pre and posttest performance suggest the treatment raised student posttest performance across nearly the entire range of student pretest performance levels.

Table 2. ANOVA of math score differences between treatment and control groups

Variable	***Estimate***	***SE***	***df***	***t***	***p***
(Intercept)	211.16	0.79	520.30	267.17	**0.000*****
Treatment	3.82	2.00	520.30	1.90	

0.057 TestTime=PostTest 3.60 0.43 448.00 8.30 0.000*** Treatment x TestTime 3.21 1.10 448.00 2.92 0.004**

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Reading Learning Gains. To ensure the observed differences in math performance among treatment and control participants were not driven by selection bias or an overall “mentoring effect,” similar analysis was conducted using nonequivalent groups design comparing students’ reading test score gains [28]. Unlike what we saw for math MAPs scores, on average students in both the control and treatment group showed similar growth in reading over the span of a year (2.1 RIT score points for the control group and 2.7 RIT score points for the treatment group). As we saw before, this growth is below the national average (4.5 RIT points).

Using statistical analyses similar to that used with math scores, we saw an overall positive growth in reading scores from pre to posttest, $\beta = 2.05$, t (417) = 4.20, $p < .0001$, but no overall effect of treatment, $\beta = 1.43$, t (492.02) = 0.67, $p = .051$, nor evidence of an interaction between the treatment group and time of the test, $\beta = 0.61$, t (417) = 0.50, $p = .619$. Overall, these findings suggest that there is no overall “mentoring effect”, or improvement in student performance due to the noncognitive effects of mentoring (i.e., social-motivational supports, relationship building) that extends beyond the targeted math learning gains. In addition, the results support the exclusion of a selection effect or the possibility that the RtL treatment students are generally better, more motivated students. The differences between control and treatment groups on math MAPs scores are likely due to the RtL treatment.

EdTech Usage and Outcomes. We investigated the role of the computer tutoring element of the PL^{2}approach to combined human-computer mentorship by analyzing the relationship between students’ EdTech usage during the program and their MAP scores. We combined data from both EdTech sources and used a measure common to both: total time in the program. Fifty-four students used MATHia and 16 used ALEKS. On average, students spent a median 102 minutes in the educational technology system during the RtL program ($M = 150.70$, S$D = 201.00$). We used a multiple regression model predicting student pre-post change, using pretest, EdTech usage (mins) and their interaction, as well as the type of EdTech used and its interaction with EdTech usage as predictors. The type of EdTech used did not have a significant relationship with score growth, $\beta = 3.16$, t (64) = 0.72, $p = .48$, and did not vary depending on amount of EdTech usage, $\beta = 0.20$, t (64) = -1.10, $p = .28$. Pretest scores also did not have a significant relationship with score growth either, $\beta = 0.03$, t (64) = -0.32, $p = .75$. Importantly, there was a positive relationship between minutes spent on EdTech and score growth, , $\beta = 0.18$, t (64) = 2.22, $p = .03$, indicating that spending more time on EdTech during the RtL program was associated with higher learning growth. Moreover, the relationship between EdTech usage and learning growth was moderated by pretest scores, $\beta = -0.001$, t (64) = -2.10, $p = .04$.

5 Discussion and Conclusion

The combined human-computer personalized approach of PL^{2}is based on the following hypotheses, which we posit as explanations for the demonstrated enhanced learning gains of the RtL participants in this study. Many marginalized students are not given sufficient learning opportunities [24]; thus, they do not get the deliberate practice they need to achieve success [5]. Educational technologies can provide such deliberate practice, which is one piece of PL^2, but only if the technology is used. Thus, the second piece of PL^{2}is the notion that human mentors provide needed social motivational support that help students engage in rigorous deliberate practice. After all, deliberate practice is motivationally challenging [4]. PL^2 helps human mentors not only personalize their math content tutoring, but also personalize motivational support. It gives mentors strategies for relationship building, which is foundational to student learning outcomes [18] and helps them personalize whether a student's effort could be enhanced by one or another motivational intervention, including growth mindset [32], valuing utility value of STEM [10], and self-efficacy building [26]. There is limited flexibility in schools to add extra learning opportunities and there is good evidence that out-of-school learning opportunities are a major source of opportunity gaps. For example, evidence of "summer slide" indicates that racial achievement gaps widen over the summer and implicate greater learning opportunities for privileged students than for marginalized students [1]. Recently, the "COVID slide" has exacerbated such opportunity gaps among marginalized students, particularly in math [15]. No amount of improvement during the school day will address this opportunity gap.

Given our goal to increase student opportunities beyond those available schools, we did not seek or create a control group that was matched for opportunities. Some students in the matched control may have attended other out-of-school programs. To be sure, out-of-school opportunities intended to support academic learning do not necessarily do so. Our results demonstrate that our out-of-school activities, which mix human and computer tutoring, do enhance student learning and quite substantially.

In addition, we hypothesize one of the reasons for the positive academic impact evidenced from the PL^{2}approach comes from the intentionally designed culturally relevant training and tutoring practices within the RtL program format. Guryan et al. [8] reports similar success of "high impact" tutoring, however, occurring during the school day taking away from instructional time for other content. In two randomized control trials (RCTs), Guryan et al. [8] states increased math scores of 0.16 and 0.37 standard deviations, respectively, with evidence of impacts existing over time. Our results indicate a larger effect size (d = 0.40), however without the rigor that comes with RCTs. Guryan et al. [8] reports a cost per participant per year of \$3,500 to \$4,300. Our research team's calculations suggest that marginal costs on the order of \$500 per student appear feasible within a few years, perhaps yielding stronger academic impact without sacrificing valued school time.

While RCTs are the gold standard experimental research method, they are not al ways practical. Quasi-experimental methods are especially valuable early in a project lifecycle to determine whether the costs of a full RCT are justified. We illustrate cost effective use of two quasi-experimental methods, propensity score matching [9, 27] and a nonequivalent dependent variables (NEDV) design [28]. Propensity score matching removes the costs of random assignment and can be straightforwardly em ployed when school partners can provide student-level demographic data. Similarly, school partnership can make employing the NEDV design simple when the school can provide two kinds of test results: one test that is aligned with the content of your instructional intervention (a math test in our case) and one test that is not (a reading test in our case).

While we presented evidence for the benefits of EdTech learning opportunities, we were not able to similarly investigate the role of the mentor. We hope to explore whether students learn more if they have mentors that provide more learning opportu nities or that use PL^2 more often or more effectively. We are also interested in deter mining the role mentor and mentee matching based on demographics and socioeco nomics plays in learning gains. Further research assessing the impact of the online Ready to Learn (RtL) program will be analyzed to determine the impact of the PL^2 system in a virtual environment with hope of increasing scalability. A fully online version was developed and implemented during the 2020-2021 school year with a RtL virtual session format consisting of virtual personalized instruction in conjunction with student self-directed use of ALEKS in a 4:1 student to mentor ratio.

Our results demonstrate progress in human-computer teaming in mentoring and tu toring providing needed out-of-school learning opportunities and producing substan tial and significant learning. More generally, this work supports the idea that greater educational equity can be achieved at reasonable costs by supporting after-school programs with technology that improves mentoring and student learning.

References

1. Alexander, K., Pitcock, S., & Boulay, M. C. (eds.): The summer slide: What we know and can do about summer learning loss. Teachers College Press (2016).
2. Autor, D.: Skills, education, and the rise of earnings inequality among the “other 99 per cent.” Science, 344(6186), 843-851 (2014).
3. Cook, P., Dodge, K. Farkas, G., Fryer, R., Guryan, J., Ludwig, J., Mayer, S., Pollack, H. & Steinberg, L.: Not too late: Improving academic outcomes for disadvantaged youth. Northwestern University Institute for Policy Research. Working Paper No. 15-01. (2015).
4. Duckworth, A. L., Kirby, T. A., Tsukayama, E., Berstein, H., & Ericsson, K. A.: Deliber ate practice spells success: Why grittier competitors triumph at the National Spelling Bee. Social psychological and personality science, 2(2), 174-181 (2011).

5. Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C.: The role of deliberate practice in the acquisition of expert performance. Psychological review, 100(3), 363 (1993). 6. Graesser, A. C., Forsyth, C. M., & Lehman, B. A.: Two heads may be better than one: Learning from computer agents in conversational trialogues. Teachers College Board. 119(3), 1-20 (2017).

7. Guryan, J., Christenson, S., Claessens, A., Engel, M., Lai, I., Ludwig, J., & Turner, M. C.: The effect of mentoring on school attendance and academic outcomes: A randomized

evaluation of the Check & Connect Program. Institute for Policy Research Working Paper Series, WP-16-18. Northwestern University (2017).

8. Guryan, J., Ludwig, J., Bhatt, M. P., Cook, P. J., Davis, J. M., Dodge, K., ... & Steinberg, L.: Not too late: Improving academic outcomes among adolescents (No. w28531). Nation al Bureau of Economic Research (2021).
9. Hansen, B. B.: Full matching in an observational study of coaching for the SAT. Journal of the American Statistical Association 99(1), 609–619 (2004).
10. Harackiewicz, J., Rozek, C., Hulleman, C, & Hyde, J.: Helping parents to motivate adoles cents in mathematics and science: An experimental test of a utility-value intervention. Psychological Science 23(8), 899-906 (2012).
11. Heller, S. B., Shah, A. K., Guryan, J., Ludwig, J., Mullainathan, S., & Pollack, H. A.: Thinking, fast and slow? Some field experiments to reduce crime and dropout in Chicago. The Quarterly Journal of Economics 132(1), 1-54 (2017).
12. Hutson, K. M.: Missing faces: Making the case for equitable student representation in ad vanced middle school courses. In Promoting Positive Learning Experiences in Middle School Education (pp. 200-216). IGI Global (2021).
13. Kim, K. H., & Zabelina, D.: Cultural bias in assessment: Can creativity assessment help? The International Journal of Critical Pedagogy 6(2), (2015).
14. Kumar, R., Rosé, C. P., Wang, Y. C., Joshi, M., & Robinson, A.: Tutorial dialogue as adaptive collaborative learning support. Frontiers in artificial intelligence and applications 158, 383. (2007).
15. Lewis, K., & Kuhfeld, M.: Learning during COVID-19: An update on student achievement and growth at the start of the 2021-22 school year. NWEA (2021).
16. Lewis, K., Kuhfeld, M., Ruzek, E., McEachin, A.: Learning during COVID-19: Reading and math achievement in the 2020-21 school year. NWEA (2021).
17. Matayoshi, J., Uzun, H., & Cosyn, E.: Studying retrieval practice in an intelligent tutoring system. In Proceedings of the Seventh ACM Conference on Learning@ Scale (pp. 51-62). Springer (August, 2020).

18. Milner, H. R.: Start where you are, but don't stay there: Understanding diversity, oppor tunity gaps, and teaching in today's classrooms (2nd ed.). Harvard Education Press (2020). 19. Muralidharan, K., Singh, A., & Ganimian, A. J.: Disrupting education? Experimental evi dence on technology-aided instruction in India. American Economic Review 109(4), 1426- 60 (2019). 20. NWEA. 2020 MAP Growth Normative Data Overview (March, 2020). https://teach.mapnwea.org/impl/MAPGrowthNormativeDataOverview.pdf 21. Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R.: Effectiveness of cognitive tutor algebra I at scale. Educational Evaluation and Policy Analysis 36(2), 127-144 (2014). 22. Popham, W. J.: Assessment bias: How to banish it. Routledge (2006). 23. Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A.: Cognitive Tutor: Applied re search in mathematics education. Psychonomic Bulletin & Review, 14(2), 249–255 (2007). https://doi.org/10.3758/BF03194060

24. Schaldenbrand, P., Lobczowski, N. G., Richey, J. E., Gupta, S., McLaughlin, E. A., Adeni ran, A., & Koedinger, K. R.: Computer-supported human mentoring for personalized and equitable math learning. In International Conference on Artificial Intelligence in Education (pp. 308-313). Springer, Cham (2021).
25. Setren, E. M.: Essays on the economics of education. [Unpublished doctoral dissertation]. Massachusetts Institute of Technology (2017).

26. Siegle, D., & McCoach, D. B.: Increasing student mathematics self-efficacy through teacher training. Journal of Advanced Academics 18(2), 278-312. (2007). 27. Stuart, E. A., &

Green, K. M.: Using full matching to estimate causal effects in nonexper imental studies: examining the relationship between adolescent marijuana use and adult outcomes. Developmental psychology, 44(2), 395–406. (2008). https://doi.org/10.1037/0012-1649.44.2.395

28. Trochim, W. M., & Donnelly, J. P.: The research methods knowledge base (3rd ed.). Cen gage (2007).

29. Van Buuren, S., & Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in R. Journal of statistical software 45, 1-67 (2011).

30. Walker, E.: Automated adaptive support for peer tutoring. Doctoral dissertation, Carnegie Mellon University (2011).

31. Walton, G. M., & Cohen, G. L.: A brief social-belonging intervention improves academic and health outcomes of minority students. Science 331(6023), 1447-1451 (2011). 32. Yeager, D. S., & Dweck, C. S.: Mindsets that promote resilience: When students believe that personal characteristics can be developed. Educational Psychologist 47(4), 302-314 (2012).