An AI-Enabled Learning System With Personalized Learning Pathways a Pilot Study of Its Impact on Learning of Statistics

Poh Nguk Lau, Temasek Polytechnic, Singapore
Steven Chee Kuen Ng, Temasek Polytechnic, Singapore
Li Fern Tan, Temasek Polytechnic, Singapore

The Asian Conference on Education 2023
Official Conference Proceedings

Abstract
AI-enabled systems offering personalized learning pathways or options are gaining imminence, showing immense potential to meet diverse learners' needs on a more practical scale. In this work, we piloted a learning resource that offers personalized learning pathways (or LeaP), powered by AI technology. The efficacy of the learning tool was evaluated using a skills test in a freshman statistics course. The results largely replicated what was found in the literature. For learners who used the resource, levels of engagement were not dependent on prior ability measured by past-semester GPA performance. The greatest difference in test scores was seen in the test task which the LeaP unit modelled after, with significant differences between learners who engaged with LeaP deeply versus those who did not attempt the unit at all. At-risk learners had poorer engagement levels and test performance compared to non-at-risk peers, which warrants a closer look at how intelligent tutoring systems (ITS) should be designed to meet their needs in online learning environments. Suggestions for future implementation and research were also proposed.

Keywords: Artificial Intelligence (AI), Applied Computing, Foundation Statistics, Human-Computer Interface, Interactive Learning Environment, Intelligent Tutoring Systems, Personalized Learning Pathways

iafor

The International Academic Forum
www.iafor.org

1. Introduction

The COVID-19 pandemic has left an indelible mark on the delivery of traditional education on a global scale. The rapid rise of Artificial Intelligence (AI) technology is gaining traction in the delivery of a personalized learning experience that adapts to the specific needs of the learners (Pantelimon et al., 2021). Its broad application in educational settings has also been driven by technological advances, resulting in improved productivity and efficiency (Chassignol et al., 2018).

1.1 ITS and AI

Intelligent tutoring systems (ITS) and AI are the two key pillars that support just-in-time, adaptive learning in an online learning space. ITS can be seen as an antecedent to AI technology; both aim to inform instructors of learners' behaviors and interactions in an online learning environment. As described by Freedman et al. (2000), ITS is a computer system that provides personalized instruction and feedback without the intervention of a human tutor. ITS is a powerful educational tool that can be customized and integrated into learning systems and relies on computer programming to enhance the learning experience with tailored lessons. With the advent of technology, ITS has evolved from the traditional, rigid computer-aided instruction models associated with hard-coded links and has been under development for decades as Graesser et al. (2018) noted in their research. With ITS, customized instruction may be delivered to diverse learners at scale, achieving the benefits and effects of one-on-one coaching to improve academic performance (Bloom, 1984).

AI can be seen as the undergirding technology to augment ITS functions. Unlike ITS, where the learning paths are more or less hard-coded and static, AI revs up the backend technology to enhance the linkage and communication between the learner and the learning content by optimizing the four main components of the learning system: the domain knowledge, learner's current knowledge level, pedagogical or instructional measures (such as assessment) and the user interface (Pappas & Drigas, 2016; Pipitone et al., 2012). The integration and synchronization of these components is the cornerstone of AI-enabled tutoring systems and is pivotal in delivering customized instruction to meet diverse learners' needs.

1.2 Learning Outcomes and Engagement in Adaptive Learning Environments

ITS, whether AI-enabled or otherwise, have been experimented with in the teaching and learning of STEM subjects such as mathematics (Bang et al., 2023; Bartelet et al., 2016; Beal et al., 2010; Eryılmaz & Adabashi, 2020) and even in non-STEM domains such as sports and dance or in business courses (Ashwin et al., 2023; White, 2020). Positive learning outcomes were noted in terms of improvement in assessment scores or user satisfaction levels (Bang et al., 2023; Bartelet et al., 2016; Beal et al., 2010, Eryılmaz & Adabashi, 2020; White, 2020).

For at-risk learners, defined as those with poor academic performance and more likely to drop out of the course (Repetto, 2018, p. 163), an online learning environment seemingly affords a higher level of learner autonomy, control, and pace to their advantage. With AI-enabled self-study tools, some work reported positive outcomes with low attainment students. For example, Bang et al. (2023) reported that the use of an adaptive learning app for mathematics produced the greatest learning gains in assessment tasks for learners from at-risk socio-economic groups. Bartelet et al. (2016) also reported their mathematics ITS optional homework tool produced pre-post gains in all levels of learning ability, with the low attainment learners benefitting more compared to the middle and high-level learners. The effects of more practice in the ITS environment (or interaction) on test gains were seen mainly in easier homework topics. Similar, Beal et al. (2010) found that learning gains were the most prominent in learners with the weakest initial mathematics aptitude, and learners who did more ITS sessions improved more than learners who accessed less of it.

1.3 Research Questions

Against the research, the current study aims to answer the following research questions (RQs):

1. What are the differences, if any, in learner engagement of an AI-enabled tool that offers personalized learning pathways (thereafter called LeaP), amongst learners of varying aptitudes in a freshman statistics course?
2. What is the impact of learner engagement in the LeaP tool on a statistics skills test?
3. How did the academically challenged (or at-risk) learners engage with the tool and what was the impact on the skills test?

2. Method

2.1 Participants and Lesson Deployment

The LeaP lesson unit was implemented in the fall semester of a foundational statistics course in October 2022. It was used as a revision package prior to a skills-based assessment, which required learners to compute and interpret various regression coefficient estimates using Excel data-analysis tool pack or the R software. The unit was embedded in a learning management system (LMS) and was accessible to learners in the second week of November, leading up to the assessment in the third week. The unit remained available until the mid-semester term test as it was expected that some learners might use the tool for term test revision. To evaluate the impact of the tool, we removed late attempts submitted after the skills assessment week. The sample size decreased from an initial enrolment level of 484 to 357 learners (see the section on “Variables and data analysis”). No ethics approval was needed as the unit was presented to all learners as a learning resource.

2.2 Configuration of the LeaP Unit

LeaP is an off-the-shelf adaptive learning tool that is integrated within the LMS. It has three main components: a mapping engine, a ranking engine and a recommendation engine. The mapping engine receives content assets and resources such as learning outcomes, learning materials (lecture notes in word or presentation files, video lectures), question pool and learner’s quiz results. The instructor needs to set up a .csv file format to capture the learning outcomes and question pool in pure text format, while ensuring a tight semantic association with the learning materials presented in the LMS. The backend semantic algorithm then ranks the extent of relatedness or relevancy between all the input assets. Lastly, the recommendation engine would combine the content relevancy and learners’ quiz performance to propose the most optimal materials to close learning gaps. LeaP is adaptive in nature as content relevancy would change each time a learner repeats a test question, ensuring that already-learned materials are not recommended.

The LeaP lesson pathway as experienced by the learner is summarized in Figure 1. Figures 3 to 6 display the learner view in the LMS interface at various stages of the learning pathway.

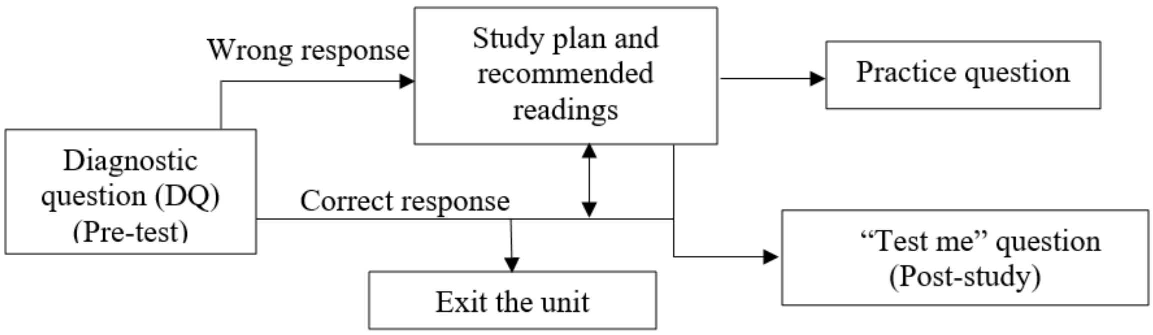

Figure 1: The learning journey in LeaP.

The test bank for the diagnostic question (DQ), practice and post-study questions consisted of 20 MCQs based on raw data from four datasets. Questions in the test bank mirrored the actual test task. The interface randomly presents one MCQ each time the learning unit is launched, whether at the DQ, practice question or post-study stage (a maximum of three DQs could be posed for a higher workload). The trigger to the recommended learning pathway is learner's performance on the DQ, taken as an indication if the learning outcome was achieved. All interaction buttons are displayed within the same browser window so that the learner has full control in the whole process, whether to continue with the recommended resources, or proceed to take a practice or post-test question. A "thumbs up" or "thumbs down" button is also available for learners to recommend a learning resource depending on the perceived utility. This allows the AI backend to adjust the relevancy of the material. An example of a DQ is shown in Figure 2. Feedback is shown upon answer submission.

Download the "Student grades" dataset. Using Excel or R/R-studio, (i) Determine the equation of the multiple linear regression trend line to predict English marks from Writing and Reading marks. (ii) What is the R^2 value? (iii) What does R^2 tell you?

Figure 2: Example of a DQ.

Figure 3: The initial landing page with the DQ or pre-test question.

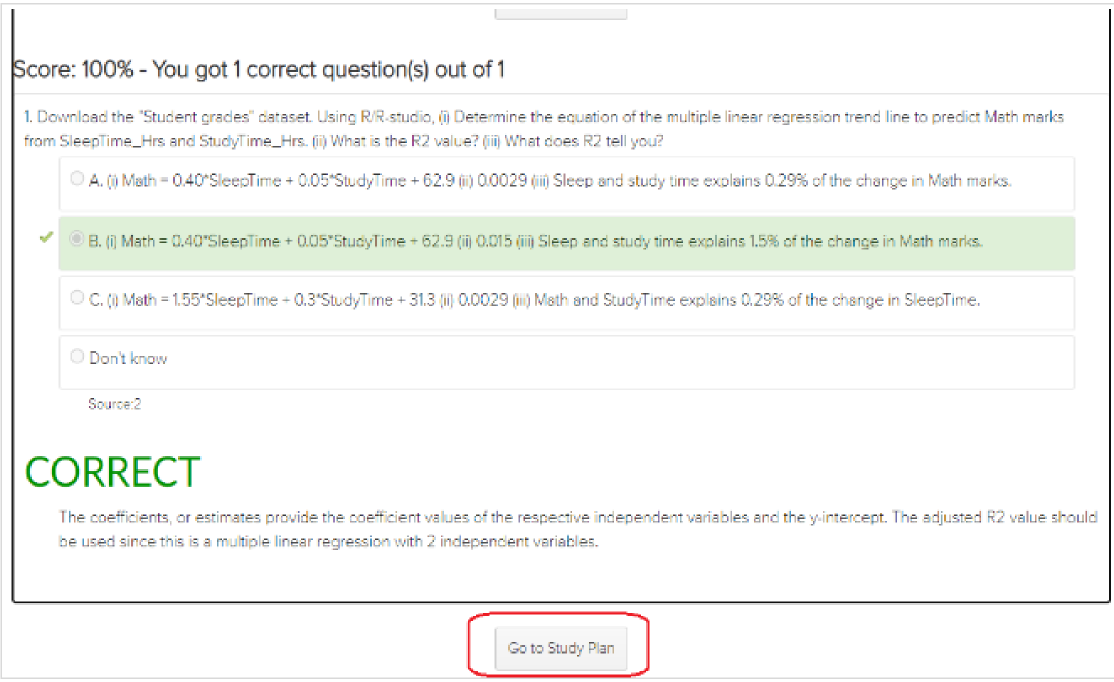

Figure 4: Correct response to the DQ, with the option to exit or continue with study.

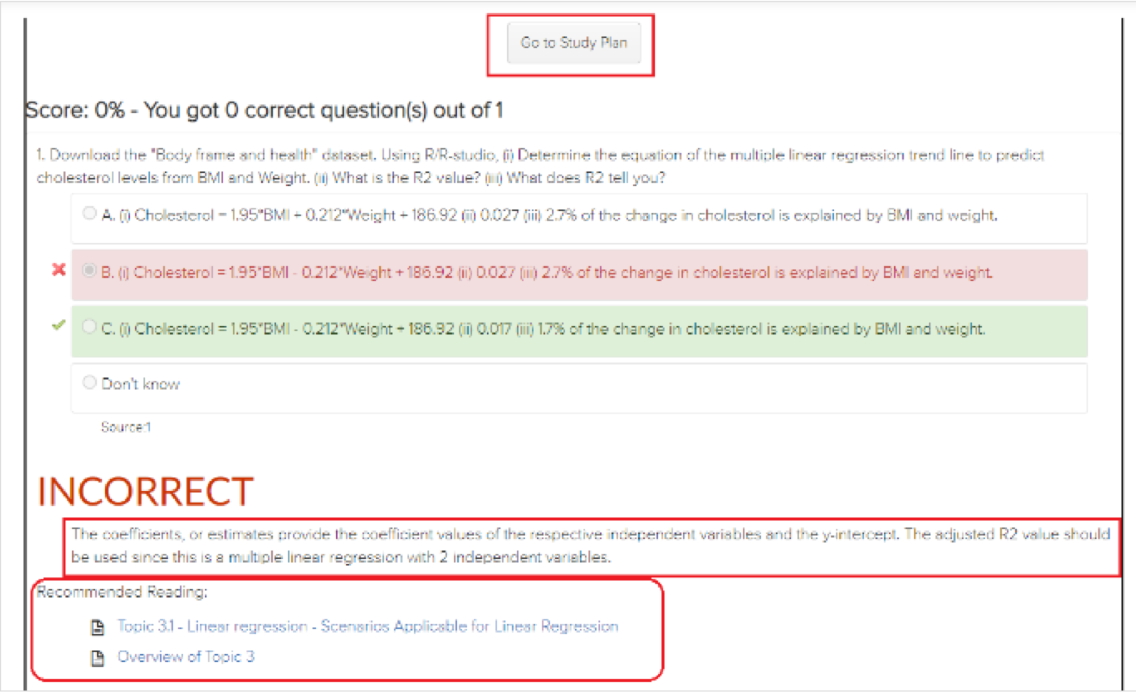

Figure 5: Incorrect response to the DQ with feedback, recommended revision resources.

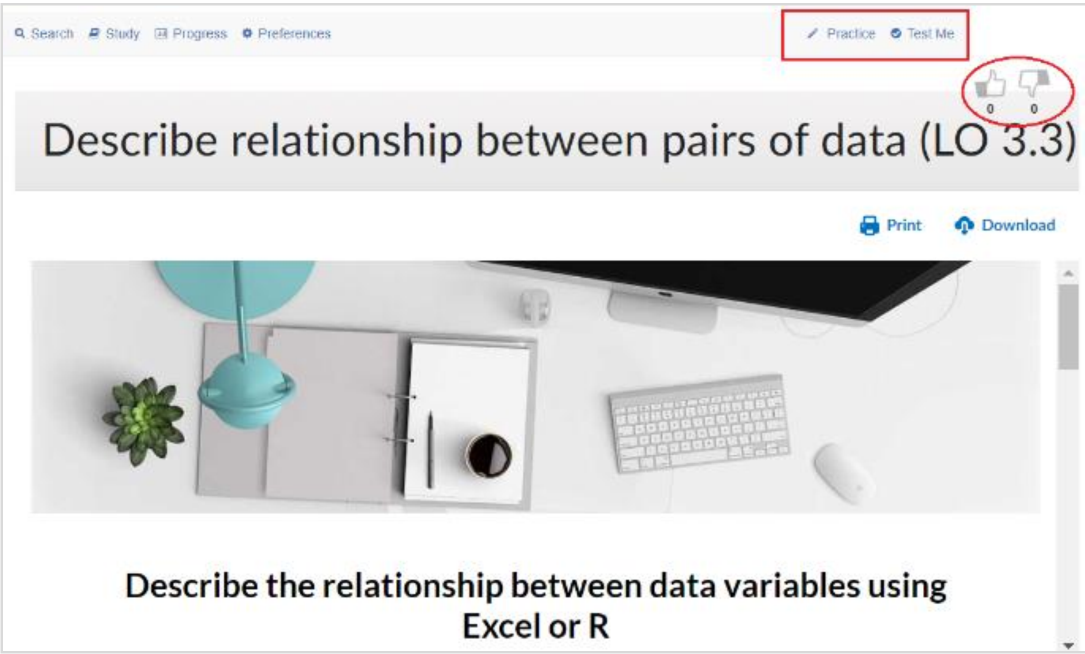

Figure 6: A study path in-progress, with the marked-out boxes showing the "thumbs-up/thumbs-down" buttons, practice and post-test question.

2.3 Variables and Data Analysis

The learner interaction data and the scores of the skills test were downloaded from the LMS. The number of trials or attempts in the pre-test, post-test and practice-test questions was used to define the LeaP engagement levels, as shown in Table 1.

Level	Description
0	Did not do the LeaP lesson at all
1	At least one attempt on pre-test but no post-test and practice question
2	At least one attempt each of pre-test and practice test but no post-test
3	At least one attempt each of pre-test and post-test but no practice test
4	At least one attempt each of pre-test, post-test and practice test

Table 1: Levels of LeaP engagement.

Prior ability was operationalized as previous semester GPA, with a maximum possible best of 4.0. The skills test comprised three questions out of a total of 20 points with one question similar to the LeaP pool. The other two questions tested concepts related to normality tests and correlation analysis. The scores were re-based to a total of 100%. Upon removing late LeaP attempts submitted after the skills test, we obtained a total of sample size of 357 learners (223 LeaP submissions and 134 non-attempts). There were 15 students who repeated the course, whose GPAs were unknown and thus excluded from GPA analysis where needed. The repeat learners were included in the academically-challenged pool for RQ3. SPSS Version 26 was used for all data analysis.

As the test scores (dependent variable) were non-normal, non-parametric tests such as the Kruskal-Wallis H-test and chi-square test were used for group differences. GPA was segregated into three levels, Low (0 to ≤2), Mid (2 to ≤3) and High (3 to ≤4). Where the assumption of a minimum expected cell counts of 5 was not fully met, the likelihood ratio was used as the chi-square statistic (Field, 2013, p. 724). The default significance level of .05 was used, except when multiple comparisons were made to illuminate paired differences. Effect sizes for paired comparison is calculated by dividing the standardized Z-statistic by the square root of the total sample size (Pallant, 2016, p. 233).

3. Results

3.1 Relationship Between LeaP Engagement and GPA Levels

Table 2 shows the GPA bands distributed across the engagement levels, excluding the 15 repeat learners. GPA percentages in each level of engagement is shown in Figure 4.

GPA	Level 0	Level 1	Level 2	Level 3	Level 4	**Total**
Low	19	2	0	3	1	**25**
Mid	60	13	7	22	8	**110**
High	46	26	13	91	31	**207**
Total	**125**	**41**	**20**	**116**	**40**	**342**

Table 2: Levels of LeaP engagement and GPA

A chi-square analysis revealed a significant relationship in the levels of engagement and GPA levels ($\chi 2$ (8, 342) = 56.9, p = .00). As shown in Figure 7, the proportion of learners in the Low and Mid GPA bands decreased as LeaP engagement levels increased. The Cramér's V value was .28, corresponding to a moderate effect size (Cramér's V, n.d.). If the no-LeaP learners were excluded, the chi-square analysis did not show a significant relationship between LeaP engagement levels and GPA ($\chi 2$ (6, 217) = 6.24, p = .40).

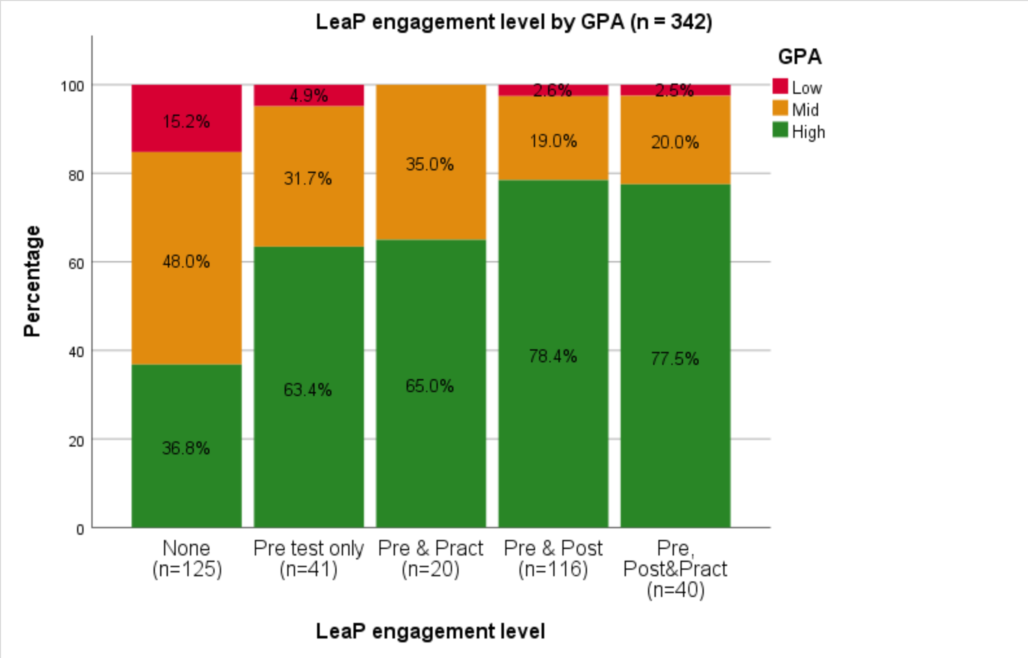

Figure 7: The proportion of GPA in each engagement level.

3.2 Impact of LeaP Engagement on Test Scores

Table 3 shows the mean, median, standard deviation and the Kruskal-Wallis H-test statistic of scores for each test question by engagement levels. As the level of engagement deepens, the score in Question 3 (which the LeaP unit models after) increased.

	0 (None)		1 (Pre-test only)		2 (Pre and Practice)		3 (Pre and Post-test)		4 (Pre, Post and Practice)		Kruskal-Wallis χ^2 (4,357)	
	M (SD)	Md	M (SD)	Md	M (SD)	Md	M (SD)	Md	M (SD)	Md		p-value
Q1	45.58 (38.83)	50	45.64 (37.42)	50	60.83 (32.23)	66.67	64.41 (36.29)	70.83	64.43 (36.61)	83.33	23.36	.000
Q2	38.06 (38.50)	50	50.57 (37.54)	50	62.50 (35.82)	50	64.30 (38.23)	50	67.07 (33.74)	50	35.31	.000
Q3	28.21 (31.47)	20	46.36 (34.96)	45	53.25 (39.94)	65	67.58 (28.52)	70	79.15 (22.10)	85	96.78	.000

Table 3: Descriptive statistics of test scores by question.
M=mean (standard deviation in brackets), Md = Median.

Figures 8 to 10 show boxplots of test scores in each question by engagement levels. As 10 pair comparisons were made, a more stringent criterion value of .005 (0.05 ÷ 10) was used. Based on this adjusted significance level, only Level 0 and 3 (none-versus-pre and post-group) had a significant difference in question 1. For question 2, besides Level 0 and 3, another pair produced significant differences in scores: the Level 0 and 4 (none-versus-pre, practice and post-test group). For question 3, the number of pairs with significant differences increased to three: Level 0 and 3 (none-versus-pre and post group), Level 0 and 4 (none-versus-pre, practice and post-test group) and Level 1 and 4 (pre-versus-pre, practice and post-test group). The effect sizes for the significant pair difference are presented in Table 4.

Question	Significant pair differences	Standard test statistic (*Z*)	Effect size
Q1	Level 0 and 3	-4.13	.22
Q2	Level 0 and 3	-5.24	.28
	Level 0 and 4	-4.13	.22
Q3	Level 0 and 3	-8.19	.43
	Level 0 and 4	-7.79	.41
	Level 1 and 4	-4.11	.22

Table 4: The effect sizes for significant pair differences per question. The effect size is calculated using the formula $Z/\sqrt{n}$, where n = 357 (Pallant, 2016, p. 233).

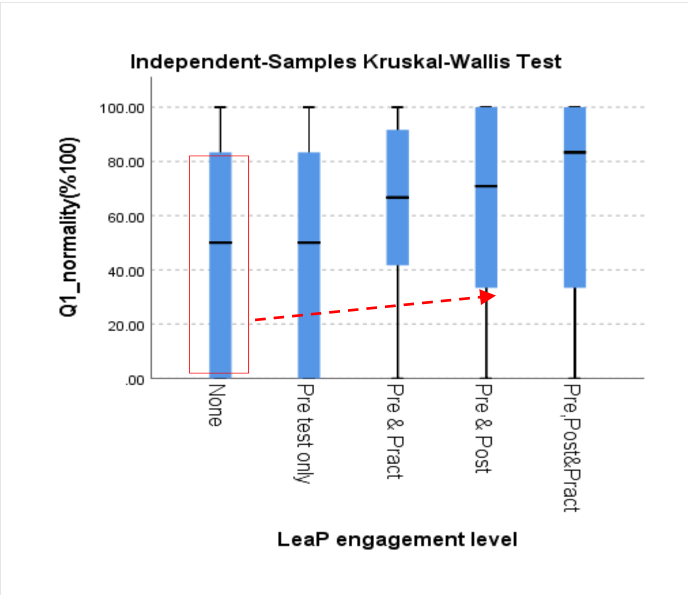

Figure 8: Boxplot for Excel test question 1. The dotted arrow represents significant differences between engagement levels. p-value = .000, adjusted for multiple comparisons.

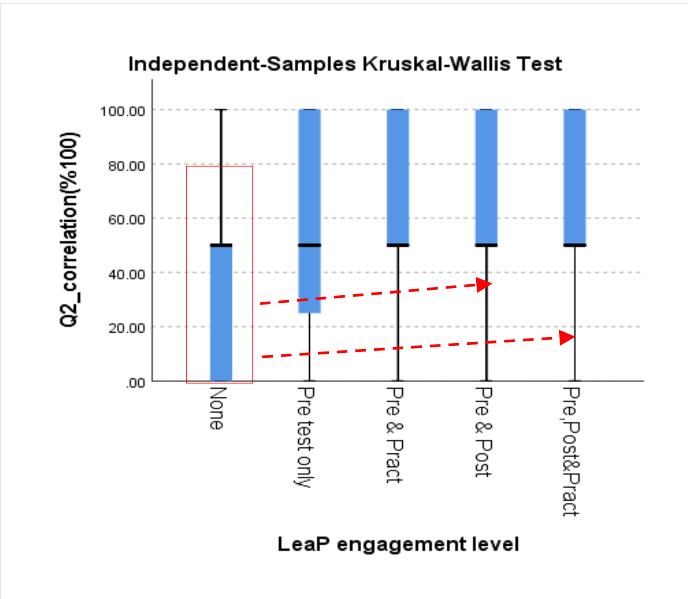

Figure 9: Boxplot for Excel test question 2. The dotted arrow represents significant differences between engagement levels. p-value = .000, adjusted for multiple comparisons.

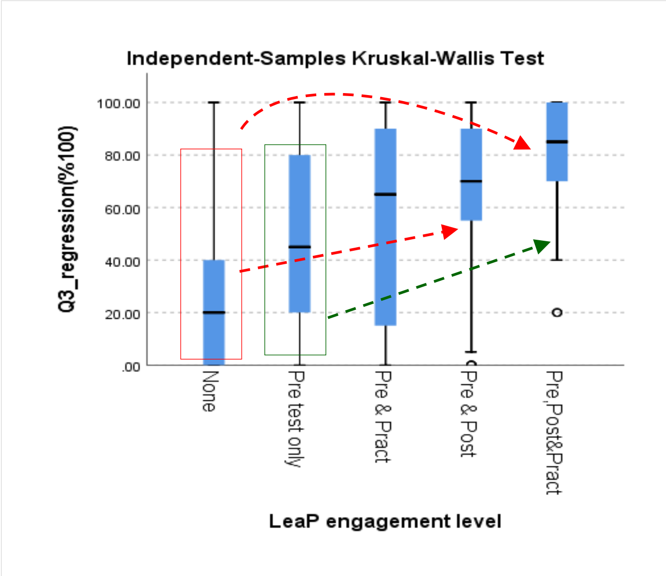

Figure 10: Boxplot for Excel test question 3. The dotted arrow represents significant differences between engagement levels. p-value = .000, adjusted for multiple comparisons.

3.3 LeaP Engagement of Academically Challenged Learners and Test Scores

For this analysis, the 15 learners (with missing GPA data) who repeated the course and the current-cohort learners with a GPA of 2 points or less were defined as at-risk. This formed a total pool of 40 at-risk learners, and 317 learners not at-risk. Table 5 summarizes the engagement counts between these two groups of learners and the mean, standard deviation and median scores. There is a significant relationship between academic status and engagement level, χ^2 (4, 357) = 25.4, p = .00. The Cramér's V effect size was .251. At-risk learners also performed significantly lower than their non-at-risk peers across all the three questions. For question 1, the Mann-Whitney U-statistic was 4,695 (Z = -2.72), p-value = .007. For question 2, U = 4,417 (Z = -3.28), p-value = .001 and for question 3, U = 3,278 (Z = -5.01), p-value = .000. As seen in Figure 11, there is a high proportion of at-risk learners who did not use the LeaP lesson unit to prepare for the skills test. Very few at-risk learners accomplished Level 4 engagement.

GPA	Level 0	Level 1	Level 2	Level 3	Level 4	**Total**	Q1		Q2		Q3	
							M (SD)	Md	M (SD)	Md	M (SD)	Md
Not at-risk	106	39	20	113	39	**317**	56.76 (37.66)	58.33	55.40 (38.63)	50.00	54.05 (35.27)	60.00
At-risk	28	5	0	5	2	**40**	39.58 (39.03)	37.50	33.75 (41.04)	0.00	24.25 (32.14)	5.00
Total	**134**	**44**	**20**	**118**	**41**	**357**						

Table 5: LeaP engagement levels and at-risk status and descriptive statistics by test question.

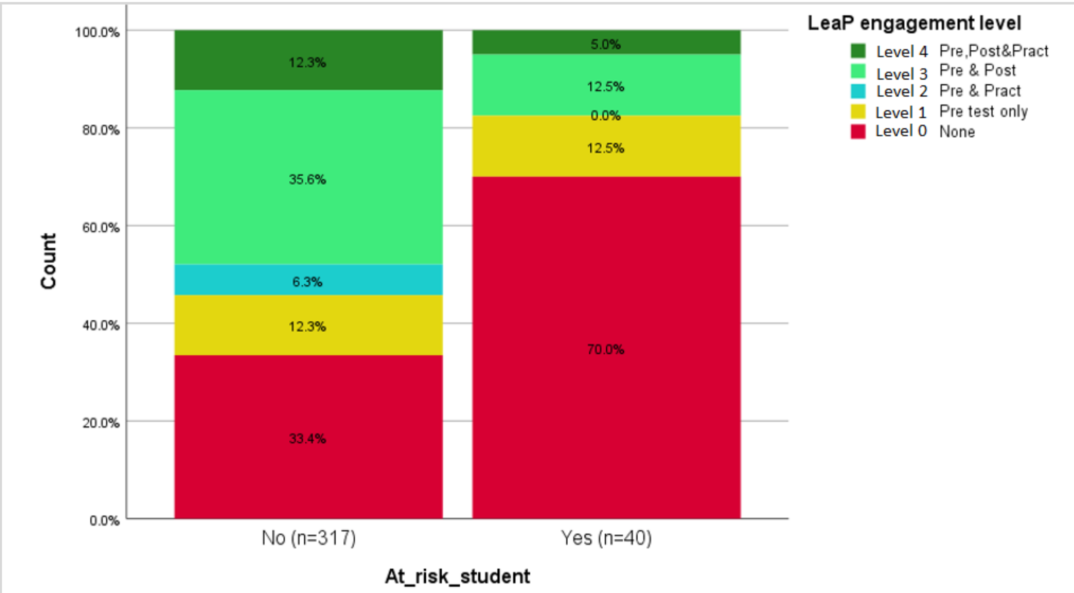

Figure 11: proportion of engagement level between the at-risk and not at-risk learners.

4. Discussion

Our first RQ was to ascertain if there is a relationship between GPA and LeaP engagement level. The results showed that about 64% of the cohort (217 learners cumulated from Level 1 to 4) had at least attempted the DQ. Amongst the 36.5% of learners who did not use the resource, about 63% of them had a GPA track record of lower than 3 points (Low and Mid categories combined), about 37% of the non-submitters being learners with a high GPA track record of 3 points and above. In contrast, for learners who at least use some LeaP, the proportion of strong-GPA attainment was much higher, reaching a maximum of 78.4% for level 3 learners with a GPA track record of at least 3 or better (see Figure 4). However, when the non-submitters are excluded, the amount of effort exerted did not differ by GPA track record, thus answering the first research question.

The results also pointed to a polarization of test performance between the non-submitters and deeply engaged LeaP learner. For question 3, the task modelled after the actual assessment question, learners who did not access LeaP at all were worse off than learners who practiced the unit beyond just doing the DQ or pre-test. In fact, the largest effect sizes in the range of approximately 0.4 were observed in question 3 between the non-doers and those learners who diligently completed the pre-test, practice, or post-test. This illuminated the second research aim in that the more conscientious LeaP learners gained better outcomes. It is important to note that low GPA learners would also benefit if they spend effort and time to use the LeaP unit. Taken altogether, the results confirmed the evidence obtained in past studies (Bartelet et al., 2016; Beal et al., 2010; Eryılmaz & Adabashi, 2020), in that practicing tasks in an ITS or a learning environment that offers personalized learning routes has alleviating effects on academic performance. Low attainment learners also benefit if they choose to engage with the resource.

The differences in engagement became very pronounced when at-risk learners were compared to non-at-risk learners. Not only did the at-risk learners performed poorly in all the three questions, a large proportion of at them (70%) did not use the LeaP tool at all, as compared to about a third of the non-at-risk students who likewise did not access LeaP. The proportion of learners engaged in deeper learning for the at-risk group was also very much lower than the corresponding levels for the non-at-risk group. There are two possible reasons for the poor levels of engagement exhibited by the at-risk learners. Firstly, due to the relatively short window of exposure (1 week), there was insufficient time for instructors to engage with learners, particularly the at-risk group (who might perhaps need more handholding and probing), to explain or communicate more clearly the goal of this instructional package. Secondly, there could be other needs troubling at-risk learners, such as family problems (Repetto, 2018, p. 165) that simply cannot be resolved by any well-intended online learning resource or system. Therefore, more work is required to understand how to support at-risk learners in a technology-based online learning space for our institutional context.

Although our study showed that learners with low GPA but engaged sufficiently with LeaP could still potentially benefit, we could not confirm any pre-post learning gains differentiated by prior ability as evidenced in the literature. This is because our outcome measure focused on a single assessment, rather than the change in score between the DQ and post-study question in the LeaP environment. While the system does generate scores on all question types, we did not choose to analyze differences because the learning scope of the LeaP module was rather narrow. Pre-post gains would be more meaningful when the environment is enriched with more content, tasks and highly varied learning pathways. Another limitation is that it could well be the case that learners simply skipped the recommended learning pathways or study plan and jumped directly to the practice or post-test questions. In other words, the better test performance may just be an effect of repeat practice, rather than a repeat study effect. Again, while the LeaP system could generate behavioral and interaction data (for example, the number of views of the recommended reading resources or study path), analyzing such data would be more meaningful with a larger content scope. However, in the light of learner choice and autonomy within a learning space offering personalized learning options, the issue of bypassing the recommended study plan may not be detrimental to learning. Akin to Bartelet et al. (2016), pre-post analysis and content engagement data could shed light on the choices and behaviors of all learners across a range of topics, including those at-risk. This could be a scope for future research.

Despite the limitations, both learners and instructors have benefitted from this simple pilot implementation. This is because classroom time was very limited, and instructors typically have very little or just no time to do revision or practice with students in class before an assessment. Based on the statistics generated by the system, we estimated that each learner spent between 1 to 2 hours self-studying the LeaP module. If we multiply this amount of time by about 500 learners, it is immediately apparent that one-on-one personalized coaching by human instructors within curriculum time is simply impractical and infeasible. We also questioned whether we could substitute the LeaP module with revision question sets or quizzes posted on the LMS, complete with automated feedback for learner's self-study. This is certainly possible, but the user interface would lack the feature that guides learners back to the lecture materials to review the concepts they faulted on. LeaP offers a more integrated interface and the advantage of recommending appropriate lesson contents to learners immediately upon answering the DQ.

5. Conclusion

In conclusion, our results replicated the main findings of previous research to support the learning efficacy of ITS or AI-enabled systems offering personalized learning options. The experience and lesson gained in this pilot trial have guided us in terms of optimizing and configuring our content to ensure that learning pathways are aligned to match learners' as-at needs. We may say with much conviction that compared to human instructors, AI-enabled personalized learning systems possess outstanding features that enable the scale-up of individualized instruction to suit the needs of diverse learners. This removes the need for tedious and constant monitoring by human instructors. We are perhaps not far away from a utopian vision of delivering personalized education to learners with reasonable effort, as Bloom (1984) hoped for.

Acknowledgements

We would like to thank the management of the School of Applied Science for supporting this research, and to Mr Martin Cai Hui and the teaching team of Statistics for Applied Science for their assistance in launching the LeaP unit.

References

Ashwin, T. S., Prakash, V., & Rajendran, R. (2023). A Systematic Review of Intelligent Tutoring Systems based on Gross Body Movement Detected using Computer Vision. *Computers and Education: Artificial Intelligence*, 100125.

Bang, H. J., Li, L., & Flynn, K. (2023). Efficacy of an Adaptive Game-Based Math Learning App to Support Personalized Learning and Improve Early Elementary School Students' Learning. *Early Childhood Education Journal*, *51*(4), 717-732.

Bartelet, D., Ghysels, J., Groot, W., Haelermans, C., & Maassen van den Brink, H. (2016). The differential effect of basic mathematics skills homework via a web-based intelligent tutoring system across achievement subgroups and mathematics domains: A randomized field experiment. *Journal of Educational Psychology*, *108*(1), 1.

Beal, C. R., Arroyo, I. M., Cohen, P. R., & Woolf, B. P. (2010). Evaluation of animal watch: An intelligent tutoring system for arithmetic and fractions. *Journal of Interactive Online Learning*, *9*(1), 64–67.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, *13*(6), 4-16.

Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial Intelligence trends in education: A narrative overview. *Procedia Computer Science*, *136*, 16–24.

Cramér's V. (n.d.). Cramér's V. https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=terms-cramrs-v

Eryılmaz, M., & Adabashi, A. (2020). Development of an intelligent tutoring system using bayesian networks and fuzzy logic for a higher student academic performance. *Applied Sciences*, *10*(19), 6638.

Field A., (2013), Discovering statistics using IBM SPSS statistics, SAGE publications.

Freedman, R., Ali, S. S., & McRoy, S. (2000). Links: what is an intelligent tutoring system? *intelligence*, *11*(3), 15-16.

Graesser, A.C., Hu, X., Nye, B.D. *et al* (2018). ElectronixTutor: an intelligent tutoring system with multiple learning resources for electronics. *International Journal of STEM Education*, *5*, 15.

Pallant, J. (2016), *SPSS Survival Manual: A step by step guide to data analysis using IBM SPSS*. 6th edn. New York, USA: McGraw-Hill.

Pantelimon, F. V., Bologa, R., Toma, A., & Posedaru, B. S. (2021). The evolution of AI-driven educational systems during the COVID-19 pandemic. *Sustainability*, *13*(23), 13501.

Pappas, M., & Drigas, A. (2016). Incorporation of artificial intelligence tutoring techniques in mathematics. *International Journal of Engineering Pedagogy*, *6*(4), 12-16.

Pipitone, A., Cannella, V., & Pirrone, R. Cognitive models and their applications in intelligent tutoring systems (2012). In G. Paviotti, P. G. Rossi., D. Zarka, D. (Eds), *Intelligent Tutoring Systems: An Overview* (pp. 59-85). Pensa Multimedia.

Repetto, J. B. (2018). Research on At-Risk Learners in K-12 Online Learning. In R. E. Ferdig & K. Kennedy (Eds.), *Handbook of research on K-12 online and blended learning* (2nd ed., pp. 163 – 180). ETC Press.

White, G. (2020). Adaptive learning technology relationship with student learning outcomes. *Journal of Information Technology Education: Research*. *19*, 113-130.

Contact emails: lau_poh_nguk@tp.edu.sg
steven_ck_ng@tp.edu.sg
li_fern_tan@tp.edu.sg