ICIC Express Letters
Part B: Applications
Volume **12**, Number **4**, April **2021**

ICIC International ©2021 ISSN 2185-2766
pp. **307–316**

M-LEARNING PLATFORM FOR ASSESSMENT AND PERSONALIZED LEARNING OF THAI LANGUAGE BY PRIMARY SCHOOL CHILDREN

Saowalak Rungrat[1], Antony Harfield[1] and Supiya Charoensiriwath[2]

[1]Department of Computer Science and Information Technology
Naresuan University
Phitsanulok 65000, Thailand
{ saowalakr60; antonyh }@nu.ac.th

[2]National Electronics and Computer Technology Center
6th Floor NSTDA Building, Rama 6 Road, Raj-thevi, Bangkok 10400, Thailand
supiya.charoensiriwath@nectec.or.th

Received September 2020; accepted December 2020

Abstract. *KidLearn is an M-learning platform for primary school children with a personalized learning component for Thai language learning. The purpose of this study is to describe and evaluate the personalization algorithm. By applying item response theory, the algorithm calculates ability in language topics based on responses to test questions and selects new content aimed at maximizing each child's improvement in ability. An experiment was undertaken in 3 schools with 47 children with low-ability or learning difficulties in Thai language reading. The results show that improvements in the children's ability in each topic were highly correlated with the ability calculated by the personalization algorithm. Therefore, as well as KidLearn providing an efficient means for boosting a child's language learning across different topics, it effectively predicts a child's language ability which provides educators an unobtrusive testing tool for monitoring progress.*
Keywords: Personalized learning, Item response theory, Intelligent tutoring systems, M-learning, Language learning, KidLearn

1. **Background.** When technology is applied to personalized learning, it should provide learners with a uniquely tailored learning path as though each learner has the attention of an individual expert. By collecting data on the learner's past activities and interactions, recommender algorithms suggest lessons, feedback, and assessments that best match the learner's ability and enable them to overcome their weaknesses [1]. Personalized learning's key benefits include a) improving learning outcomes and learning experience, b) supporting a more active approach to teaching and c) enabling learning at scale in a sustainable and cost-effective way. A key ingredient of personalized learning is a system for effectively evaluating the learner's ability and recommending suitable learning content [2].

Computerized Adaptive Testing (CAT) is used to precisely evaluate an examinee's ability by providing a tailored path through a bank of test items. A key ingredient is the algorithm that selects the most appropriate test item based on the examinee's ability [3]. Both CAT and PL share a need to effectively evaluate the ability of participants. What they differ is that CAT recommends test items with the purpose of obtaining a more accurate estimate of the ability, while PL recommends interventions with the purpose of increasing the ability.

Several researchers have noticed the similarity and adapted CAT techniques to PL. The most popular technique, Item Response Theory (IRT), is a statistical measurement model to determine a test taker's ability and their probability of answering a given question

DOI: 10.24507/icicelb.12.04.307

correctly [4]. Given sufficient assessment data (the "responses"), an IRT model is applied to obtaining the difficulty parameter and the discrimination parameter for every question (or "item"). For a given test taker, these parameters together with the responses for other questions are sufficient to predict the probability that the test taker will respond to the item correctly. The technique enables assessment systems that dynamically select items to maximize the information about the ability of the test taker and to end the test when the system can predict the test taker's answer above a given confidence threshold. IRT can significantly reduce the length of assessments by up to 50% [5]. While IRT is highly popular in commercial CAT products, examples of IRT applied to personalized learning are relatively rare and are not yet found in commercial products. The research into IRT for personalized learning can be grouped into two broad categories: assessment-focused and training-focused.

In the area of primary education, the eDia system [6] is an e-learning assessment platform covering reading, mathematics and science used in a large number of schools in Hungary for a number of years. IRT is applied to establishing formative assessments that enable diagnostics and improvements in teaching. Results of the long-term study suggest IRT and the platform supports adjusting teaching and learning processes to the individual needs of students.

Within higher education, numerous studies have utilized IRT in assessments of computer science related subjects, such as the adaptive assessment for introductory programming by Vega et al. [7]. Typically, such systems can recommend appropriate problems based on student ability. In a study by Yacob et al. [8], undergraduate students in a programming course experienced different learning paths through multiple choice problems that were adapted based on item difficulty and learner ability. The system filtered unsuitable course materials for students, and also helped identify the items most likely needing for modification by teachers. Kustiyahningsih and Cahyani conducted a study on IRT in e-learning [9] that found students who were adaptively served questions based on IRT showed a greater increase in ability than students who were served all questions.

In the second category of related work, there are several examples of IRT applied directly to training or learning. A suitable example is how IRT can be used for vocabulary practice as proposed by Chen and Chung [10] in their work on a mobile-based e-learning system for higher education students studying English as a foreign language. The proposed algorithm chooses a suitable strategy for extending or shortening the memory cycle activities based on the ability of students and the difficulty of the content.

Other examples of IRT for personalized learning tend to focus on computer science courses at university. The recent study by Maddalora [2] proposed that diagnostic assessments are administered after each engagement with the source materials and IRT can calculate the "shortest learning sequence". After each engagement, the materials are reduced by removing those materials that the student has already gained mastery. A personalized Web-based instruction system is also developed for an introductory programming course at university by Chen et al. [3] combined IRT with an existing courseware system. By taking account of the information value of each courseware, the system could match courseware with learner ability and thus deliver "personalized curriculum sequencing".

In previous work [11], the authors proposed a personalized learning system for learning English using IRT which is unique in calculating student ability across topics, as opposed to overall levels of ability as implemented by Maddalora [2], Kustiyahningsih and Cahyani [9] and Chen and Chung [10] described above. The purpose of this paper is to implement the algorithm and to test its validity and performance. Therefore, the significance of the current study is *a concrete implementation of a training-focused approach to personalized learning using IRT* and an evaluation of its effectiveness within the relatively unexplored domain of primary language education. The rest of the paper is organized into 3 sections: the Methodology section that describes the theory, implementation and testing of

KidLearn tutoring system, the Experimental Result and Discussion section that presents the outcomes of testing with children in three Thai schools, and the Conclusion section that summarizes the paper.

2. **Methodology.** The method covers preparing the algorithm and content on KidLearn; deploying KidLearn in schools and collecting responses to evaluate the algorithm; administering pre/post-tests to evaluate student abilities for comparison with the algorithm.

2.1. **Preparation of the algorithm and content.** KidLearn is an application on the iPad (designed and implemented as part of the study) for children to practice Thai language reading skills. The content in KidLearn was devised by experts in Thai language learning with a focus on letter sounds for children aged 6-8 years old. The content consists of 98 questions, divided into 6 topics based on similar sounding letters (from 42 Thai consonants [12]). The topics are ordered by experts from easy to difficult, consisting of topic A (21 questions), topic B (18 questions), topic C (21 questions), topic D (22 questions), topic E (9 questions) and topic F (7 questions). Each question can be served in 4 different formats: "letter song", "train drag-n-drop", "fruit in basket" and "alphabet balloon". Figure 1 shows two example questions. The first (on the left) is a question from topic C in the "fruit in basket" format. The child must listen to the word and drag the letter for the starting sound to the basket. The second (on the right) is from topic A in the "letter song" format, and involves pressing on a letter instead of dragging. The application first selects the easiest topic and administers 10 questions from that topic to the child. It chooses the next topic according to the ability of the child in each topic using the proposed personalized learning algorithm [7].

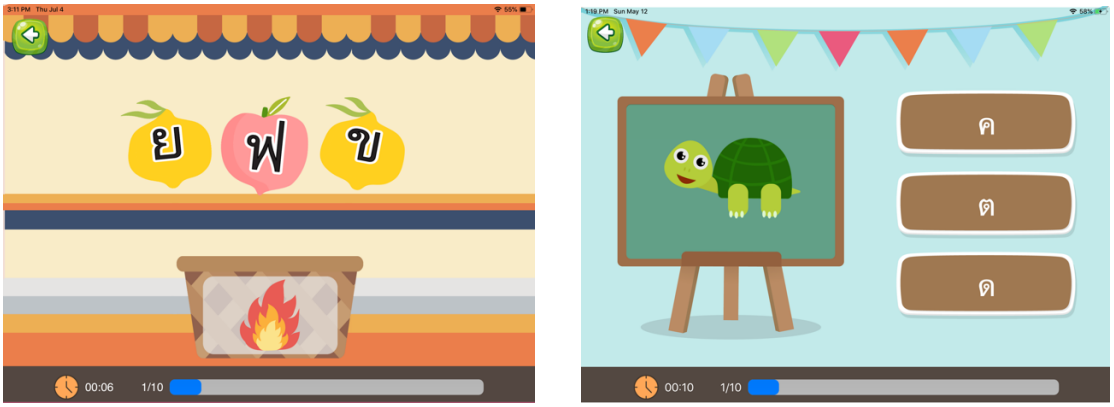

FIGURE 1. Example questions in "fruit in basket" (left) and "letter song" (right) formats

The algorithm calculates the ability of a child in a particular topic according to their responses to items in the topic and the difficulty and discrimination of each item. In this calculation, the difficulty and discrimination parameters are a measure of how useful the item is in differentiating between participants of high and low ability. The discrimination parameter and the difficulty parameter define the item response function (1) which is represented by the Item Characteristic Curve (ICC). The values of the discrimination and difficulty parameters affect the slope and the horizontal offset of the characteristic curve, respectively. A high difficulty value indicates that the item a person of higher ability is more likely to answer correctly. A high discrimination value indicates a stronger classification power. The Two-Parameter Logistic model (2PL) calculates the probability from the difficulty and discrimination of each item [4].

$$P_j(\theta) = \frac{e^{Da_j(\theta - b_j)}}{1 + e^{Da_j(\theta - b_j)}} \tag{1}$$

where $P_j(\theta)$ is the probability that the participant will give the correct response to item j, a_j is the discrimination parameter of the item, b_j is the difficulty parameter of the item and D is a constant value of 1.702. In general, calculating the estimation of each child's ability uses the Maximum Likelihood Estimation (MLE) method applied with the Newton-Raphson method to calculate the probability maximum ability of the child, as in Formula (2).

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \frac{\sum_{i=1}^{N} -a_i \left[u_i - P_i\left(\hat{\theta}_s\right)\right]}{\sum_{i=0}^{N} a_i^2 P_i\left(\hat{\theta}_s\right) Q_i\left(\hat{\theta}_s\right)} \tag{2}$$

where $\hat{\theta}_s$ is the estimated ability of the child within iteration s, a_i is the discrimination parameter of item, u_i is response for item i and N is the number of responses, $P_i\left(\hat{\theta}_s\right)$ is the probability of the correct response to item i from ICC in Equation (1) at ability level $\hat{\theta}$ within iteration s. $Q_i\left(\hat{\theta}_s\right)$ is the probability of incorrect response to item i calculated by $1 - P_i\left(\hat{\theta}_s\right)$.

In CAT, the above model is applied per test (for a specific bank of questions). The proposed algorithm applies the model across multiple topics (where each has its own questions) and therefore the overall ability of the child within the system can be obtained from Equation (3).

$$\bar{\theta}_s = \frac{\sum_{t=1}^{T} d_t \hat{\theta}_{st}}{T} \tag{3}$$

where $\bar{\theta}_s$ is the averaged estimated ability of the child across T topics and $\hat{\theta}_{st}$ is the ability in topic t within iteration s. The goal of the algorithm is to maximize $\bar{\theta}_s$ for each child. At each new iteration $s+1$, the algorithm chooses the topic t that has the most potential to increase $\bar{\theta}_s$. The parameter d_t enables topics to be weighed independently.

In KidLearn there are 6 topics and equal weighting is applied to each topic. Therefore, the algorithm chooses the topic where the child has the lowest ability after administering each iteration of 10 items as illustrated in Figure 2. The ability in a topic is calculated from all the responses in that topic – at the end of the first iteration there will be 10 items, and the next time that topic is administered there will be 20 items, and so on.

The first time the KidLearn application is used, there is no response data to calculate the difficulty and discrimination parameters – the so called "cold-start" problem [13].

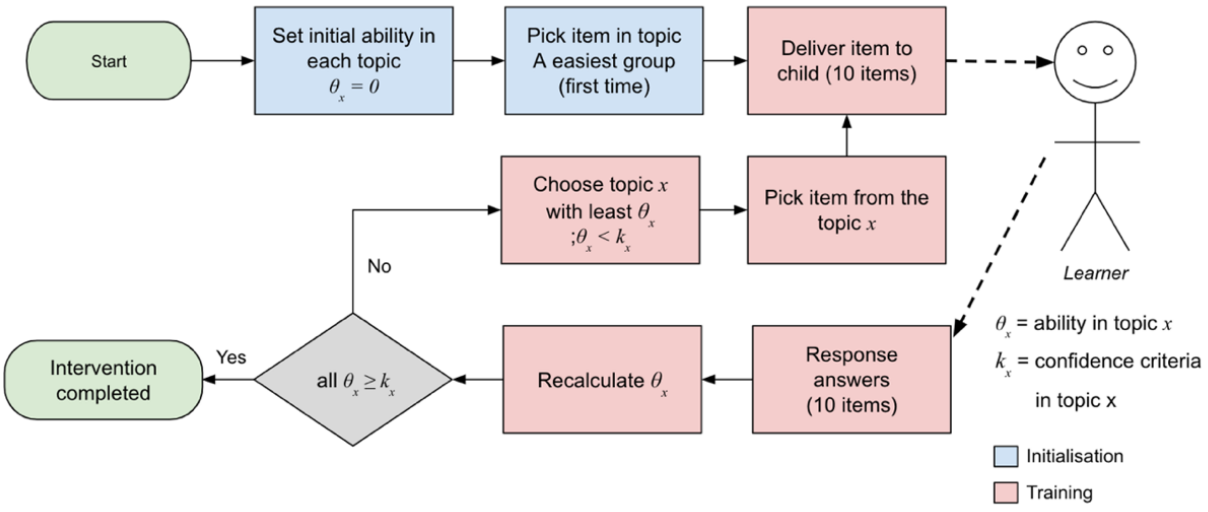

FIGURE 2. (color online) The algorithm design with IRT of system

For this case, the experts who selected the questions rated them by difficulty: easy is -3.0, medium is 1, hard is 3.0. The discrimination parameter set default value 0.5 for all items. These default values were used for the initial iteration of the experiment and were then replaced with values calculated from actual responses (when each item has enough responses to calculate the parameters with IRT). In this way, the system is able to operate in the early cold-start phase.

The KidLearn iPad application is connected to the KidLearn API hosted on a cloud server. The logic in Figure 2 for delivering items, calculating abilities and selecting topics is performed by the API. After each iteration of 10 items, the application sends the responses to the API and the algorithm recomputes the ability for that student. Furthermore, the API provides a complete history of every response to every item, and the progress of the children in terms of their ability in each topic. The data was exported for the analysis in this paper.

2.2. **Deployment in schools and data collection.** The KidLearn application was deployed at 3 primary schools where a preliminary assessment of the children's language ability had already been performed [14]. Children who scored within the 10th percentile were selected for the current study, a group considered slow learners bordering on learning difficulties. A total of 47 children were selected. The application was used by each child for 20 minutes per day (during lunch breaks) for 4 days per week (Monday to Thursday) for 4 weeks.

Each child's progress from each interaction was saved in the application and a child could come back to the same place on subsequent interactions. The algorithm always took account of all the abilities per topic $\hat{\theta}_{st}$ from previous sessions in personalizing the next iteration of questions for the interaction. KidLearn has a threshold (set by the school or experiment) for when a child can stop the activity. In this experiment, when a child's ability θ_x reached a threshold k_x in every topic x, the sessions were no longer compulsory for the student. A threshold was set from early studies at an ability of 2.0 in every topic.

2.3. **Evaluation of ability development and comparison.** Two sources were integrated to evaluate the outcomes of the study: pre-post tests taken outside of the system and responses recorded within the KidLearn system.

Each child in the study took a pre-test before the 4 week period, and a post-test at the end. The purpose of the pre-post tests is two-fold. First, it measures the improvement in ability of the child (at least partly) due to the intervention of KidLearn. Second, the correlation between the post-test scores and the improvement in ability in each topic calculated by KidLearn gives the accuracy of the system in predicting each child's ability.

For the pre-post tests, an evaluation using paired t-test with 95% confidence determines the significance of the child's improvement. The hypothesis is that scores on pre-test and post-test are significantly different. For the correlation, Spearman's rank correlation coefficient is calculated between the post-test scores and the ability as calculated by the KidLearn algorithm at the end of the intervention.

3. **Experimental Result and Discussion.** The first result is the improvement in ability as measured by the pre-post tests. Table 1 shows the improvement by topic of the 47 children that used the KidLearn system. A paired t-test indicates an improvement above 99% confidence in all topics due to a p-value of 0.003 for topic A and < 0.001 for others.

The second result is that the post-test results were highly correlated with the ability as calculated by the KidLearn algorithm. Table 2 shows the means of the post-test and abilities from KidLearn. The result implies that the ability determined by the algorithm is consistent with the actual ability of the child. The correlation is stronger when there are more responses, as seen from topics D, E & F which have fewer responses and hence a lower probability of correlation. It suggests the algorithm would have more confidence in

TABLE 1. Paired samples statistics score pre-test and post-test ($N = 47$)

Comparative issues #	**Pre-test**		**Post-test**		**T**	**P**
	$\bar{\chi}$	**S.D.**	$\bar{\chi}$	**S.D.**		
Topic A	8.073	2.114	9.049	1.303	−3.114	.003
Topic B	5.610	1.263	6.463	0.745	−5.391	.000
Topic C	8.244	2.289	9.805	1.364	−4.585	.000
Topic D	5.951	2.224	8.171	1.548	−7.040	.000
Topic E	2.610	2.084	4.634	2.022	−7.516	.000
Topic F	1.707	1.647	3.951	2.247	−8.242	.000

TABLE 2. The children's ability between actual ability from post-tests and ability from KidLearn

	Topic A	**Topic B**	**Topic C**	**Topic D**	**Topic E**	**Topic F**
Post-test mean score	8.561	6.037	9.025	7.061	3.622	2.829
Post-test number of questions	10	7	11	11	7	7
KidLearn mean ability	19.261	15.910	18.659	18.269	7.765	5.738
KidLearn number of items	21	18	21	22	9	7
KidLearn number of responses	2317	2131	2014	1207	455	326
Correlation coefficient	0.999	0.977	0.887	0.825	0.900	0.767

TABLE 3. The learning sequence of selected children

Child #	**Sequence of topics as delivered by the algorithm**
1	A → B → C → D → E → F → D → D
2	A → B → C → D → E → F → B → B → B → A → A → B → B → B → B → B
3	A → B → C → D → E → F → A → D → A → D → B → B → B → B → D → A → A
4	A → B → C → D → E → F → B → C → B → B → B → B
5	A → B → C → D → E → F → A → D → B → A → A → A → A → B → B → B
6	A → B → C → D → E → F → C → B → A → A → A → A → A → A → A → C → B
7	A → B → C → D → E → F → A → A → A → A → A → A → A
8	A → B → C → D → E → F → C → C
9	A → B → C → D → E → F → A → A → A → A → A → A → C → C → A → A → C → C → C
10	A → B → C → C → D → E → F → C → A → A → D → A → A → A → A → A

its recommendations if it set a minimum number of responses before the recommendation was enabled. As the number of topics increases this might become unfeasible and therefore additional algorithm steps could flag the topics with insufficient responses to be confident of the ability.

The third result is that the learning sequence of topics proposed by the algorithm in KidLearn is sufficient for recommendation, but there are possibilities for improving the algorithm. In Table 3, 10 children were selected from the 47 children to show learning sequence recommended by the tutoring system from Equation (3). When starting, each child has the same ability in each topic (assume a value of zero), and therefore the algorithm will select the first available topic (which is A). At the end of one intervention with items from topic A, the algorithm recalculates the child's ability for topic A: if the child performed poorly on topic A then a second intervention of topic A would follow. In most cases from Table 3, the child performed sufficiently well to obtain an ability for topic A that is above the zero level for the remaining topics, and hence at the end of topic A the algorithm selected the next available topic for the next intervention with items from topic B. Child 10 performed poorly on topic C and therefore it was repeated before moving on to topic D. After 7 interventions (covering all topics), topic C was still the weakest topic for child 10 and it was recommended twice again – the child eventually achieving

sufficient ability in topic C to move onto other topics. Similarly, and highly evident, child 7 was recommended 7 consecutive rounds of topic A in order to bring their ability level on topic A up to that of the other topics. Note that if the child has similar ability in several topics then the algorithm will pick the weakest based on the IRT calculation from their responses, which may mean that they cover a wide range of topics instead of repeating one or two – as is in case with child 2.

The results also show a large variation in the number of times that a child covered each topic. Child 2 completed twice as much material as child 1, despite each child being given the same amount of classroom time. Child 2 completed each round of questions faster, but with more errors, particularly in topics A and B. At the end of the classroom time, child 1 had a higher ability in topics A and B compared to child 2. The 9th child had the longest test sequence, undertaking topics A and C 8 times and 5 times respectively. Each intervention within the same topic is different, as items are randomly selected from a pool.

To understand the algorithm, it is helpful to examine the abilities of individual children in each topic. From Table 3, we take children 2 and 10 to plot their ability in each topic over time (where the x axis is interventions) as shown in Figure 3. In each intervention, the ability will change only for the topic that was recommended by the algorithm. After 1 intervention (of topic A), both perform positively, although child 2 performs better than child 10 (topic A ability is $\sim$0.4 versus $\sim$0.3). They both also perform positively in topic B as seen by the increase in their ability. However, (as was mentioned earlier from Table 3) child 10 performed poorly on topic C, resulting in a calculated ability for topic C of -2.5 after the first intervention. Therefore, whereas child 2's 4th intervention was topic D, child 10 was repeating topic C for their 4th intervention. In topic F they both performed positively; child 10 performs better than child 10 (topic F ability is $\sim$2.03 versus $\sim$2.52).

The abilities of all the children at the end of all interventions as shown in Figure 4 indicate above average ability (average is above zero) measured against the IRT calculation performed in the pre-experiment. The average ability in topics E and F appears exceptionally high which could be explained by inaccurate or insufficient data in the pre-experiment IRT calculations, leading to sub-optimal choices of values for difficulty and discrimination for some (maybe all) of the topics E and F. An alternative explanation is that the questions in topics E and F were easier to learn for students than experts predicted. The experts chose topics A-F in terms of difficulty, with A being the easiest and F being the hardest (with the early topics being a prerequisite for later topics). Given this information, the expectation would be that Figure 4 should be an inverse relation of topic to ability. However, the data does not appear consistent with the experts' prediction of difficulty or progression. An unintended consequence of this method is that it can be used as a validation technique for experts' selection of content for topics. Further work could be undertaken to determine if the algorithm could predict which items are "out of place" in a given topic.

4. **Conclusion.** Through the experiments undertaken on the KidLearn platform, the study concludes that KidLearn, in particular the underlying algorithm derived from IRT that is presented in Figure 2, can estimate the ability of children across multiple topics in a way that is consistent with pre-post test results. The results that showed the children's ability increased in each topic after using the KidLearn system could be a function of the quality of the content. However, the key result is that the ability calculated by the system is highly correlated with the actual ability of the children as determined by the pre-post test. There are several areas that the algorithm and the overall platform could be improved as follows.

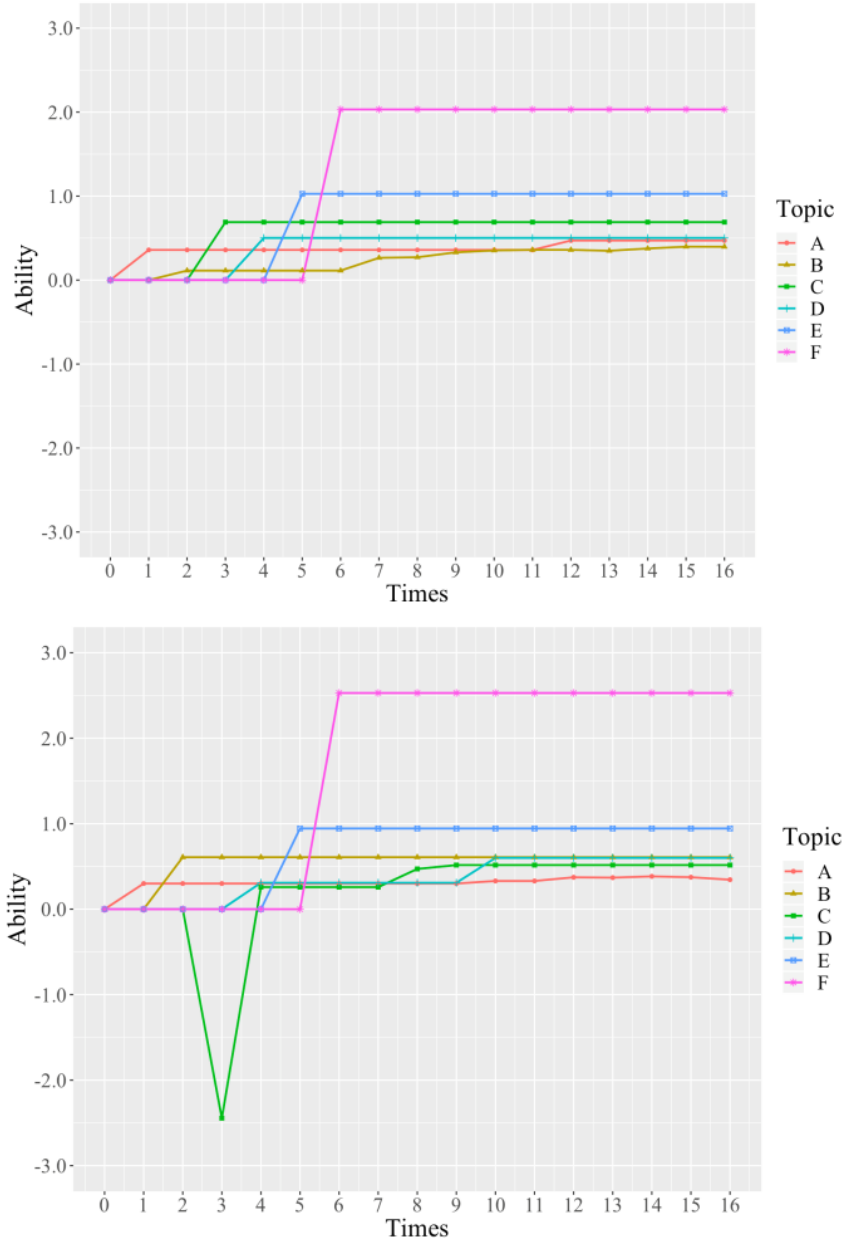

FIGURE 3. Ability progression during one session for child 2 (top) and child 10 (bottom)

Firstly, the sequence of recommended topics could benefit from some additional rules or logic. As was evident from the results, some children experienced severe repetition of topics when they were unable to achieve an ability score above the other topics. The algorithm could be modified to avoid this repetition by not selecting any topic that is repeated x times.

Secondly, topics E and F produced too high ability score which meant that they were typically only delivered for a single iteration. This was mostly caused by the cold-start approach which involved experts rating the items, and the items being easier than the experts predicted. As it turns out, the IRT approach is well-suited to detecting errors in the experts predictions. However, in the case of this experiment the cold start values chosen by experts caused irreparable damage to the children's ability scores from which the algorithm could not recover. In a future experiment it would be important to understand how to choose the initial parameters for difficulty and discrimination when the system does not have sufficient responses to calculate itself.

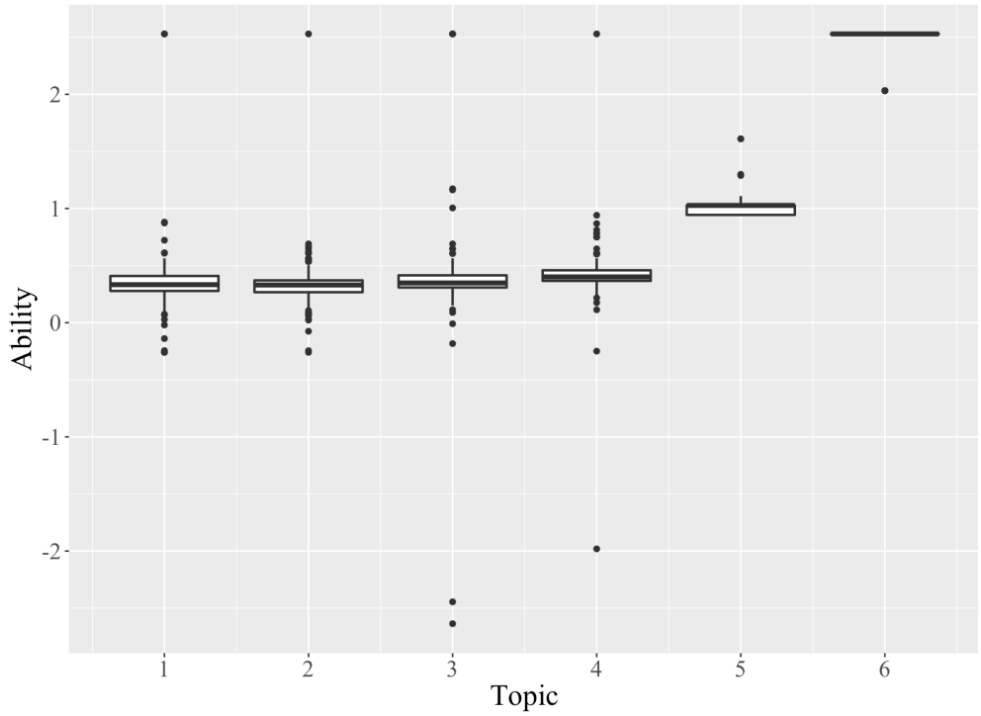

FIGURE 4. Ability for all children (as calculated by the algorithm) at the end of all interventions (box plot with outliers)

Finally, there was some bias in the results due to children with learning difficulties being chosen as the group for the experiment. A more realistic set of difficulty and discrimination parameters would be calculated from a sample of responses from the population of students instead of only learning difficulties students. If the system was deployed to the entire school (or an entire district) then the ability scores for each child would make a meaningful comparison to determine when children's learning difficulties were overcome.

Overall, the results achieved and potential problems of the KidLearn platform show potential for further research into delivering personalized learning that is based on a mathematical approach to recommending content – either in the refinement of the algorithm or in delivering different domains of learning material.

Acknowledgment. The authors would like to express thanks to Issarapa Chunsuwan and Kanokporn Vibulpatanavong for guiding the development of KidLearn, and to Samakkhi Rat Bamrung School, Pathumthani Municipality Secondary School and Pongsuwanwittaya School for testing the KidLearn tutoring system. This project was funded by Thailand Graduate Institute of Science and Technology (TGIST) programmed under National Science and Technology Development Agency (NSTDA).

REFERENCES

[1] D. Sampson and C. Karagiannidis, Personalised learning: Educational, technological and standardisation perspective, *Interactive Educational Multimedia*, vol.4, pp.24-39, 2002.

[2] M. Maddalora, Personalized learning model using item response theory, *International Journal of Recent Technology and Engineering (IJRTE)*, vol.8, pp.811-818, 2019.

[3] C. M. Chen, H. M. Lee and Y. Chen, Personalized e-learning system using item response theory, *Computers & Education*, vol.44, pp.237-255, 2005.

[4] B. Baker and H. Kim, *Item Response Theory: Parameter Estimation Techniques (2nd Ed., Rev, and Expanded)*, Taylor & Francis Group, United States, 2004.

[5] M. Yang and T. Kao, Item response theory for measurement validity, *Shanghai Arch Psychiatry*, vol.26, no.3, pp.171-177, 2014.

[6] B. Csapo and G. Molnar, Online diagnostic assessment in support of personalized teaching and learning: The eDia system, *Advancements in Technology-Based Assessment: Emerging Item Formats, Test Designs, and Data Sources*, 2019.
[7] Y. L. P. Vega, J. C. G. Bolanos and G. M. F. Nieto, Application of item response theory (IRT) for the generation of adaptive assessment in an introductory course on object-oriented programming, *Frontiers in Education Conference Proceedings*, 2012.
[8] A. Yacob, N. Hj. Ali et al., Personalized learning: An analysis using item response theory, *International Scholarly and Scientific Research & Innovation*, vol.8, no.4, pp.1107-1113, 2014.
[9] Y. Kustiyahningsih and A. Cahyani, Computerized adaptive test based on item response theory in e-learning system, *International Journal of Computer Applications*, vol.81, no.6, pp.6-11, 2013.
[10] C. M. Chen and C. J. Chung, Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle, *Computer & Education*, vol.51, pp.624-645, 2008.
[11] S. Rungrat and A. Harfield, Applying item response theory in adaptive tutoring systems for Thai language learners, *The 11th International Conference on Knowledge and Smart Technology*, pp.67-71, 2019.
[12] Thai Language Institute Bureau of Academic Affairs and Educational Standards Office of the Basic Education Commission, *A Guide to Teaching Reading and Writing to Spelling Words*, The Agricultural Cooperative Federation of Thailand, Ltd., Bangkok, 2018.
[13] B. Lika, K. Kolomvatsos and S. Hadjiefthymiades, Facing the cold start problem in recommender systems, *Expert Systems with Applications*, vol.11, pp.2065-2073, 2014.
[14] I. Chunsuwan, N. Ruangdaraganon, K. Vibulpatanavong et al., Using Kidarn application to assess Thai early reading skills: Evaluating validity and reliability, *Asia Pacific Journal of Developmental Differences*, vol.7, no.2, pp.265-280, 2020.