Reflective Dialogues and Students’ Problem Solving Ability Analysis Using Clustering

Sedigheh Abbasnasab Sardareh[1], Saeed Aghabozorgi[2] and Ashish Dutt[2]

[1]Faculty of Education, University of Malaya
[2]Faculty of Computer Science & Information Technology, University of Malaya,

abbasnasab@gmail.com

***Abstract*:** In the context of one-on-one instruction, reflective dialogues help students advance their learning and improve their problem solving ability. The effectiveness of one-on-one instruction with respect to learning through dialogue is highlighted by researchers and educators. However, little if any, is known about how reflective dialogues may lead to learning improvement and predict students’ problem solving ability. This information can be extracted from large educational datasets using data mining techniques. Consequently, this study aims at mining USNA physics dataset applying a two-level clustering approach to find patterns in the data and identify how reflective dialogues predict students’ problem solving ability. The results indicated that reflective group performed better on the hourly exams. Control subjects took lower average count of steps during problem solving activity and the average of duration was longer in control group. Also, higher average of correct answers and average count of attempts was found in reflective condition. Yet, control group had a higher level of incorrect answers as compared to reflective group. During the intervention, reflective subjects had higher level of average count of attempts and lesser average count of deletion. Reflective group asked for less hints, had lesser count of problem and requested for calculator less than control subjects. The results of the analysis help educators plan more effective tutorial dialogues in intelligent tutoring systems (ITSs).

***Keywords*:** educational data mining, clustering, reflective dialogues, problem solving ability

1. Introduction

Social constructivists such as Vygotsky emphasize the importance of learning as a social activity. Vygotsky stressed that learning is an active process in which learners interact with other people and objects in the learning context [1-2]. In Vygotsky’s viewpoint, learning context should provide learners with guided instructions so that they are able to monitor and adapt their learning through interactions.

The social constructivist movement acknowledges the importance of social interaction and more knowledgeable peers in shaping learners’ experiences [3-4]. Instructional scaffolding given by more knowledgeable peers can help learners bridge their learning gap and consequently improve their learning. One way to provide students with effective scaffolding is through one-on-one instruction. In the context of one-on-one instruction, reflective dialogues help students advance their learning and improve their problem solving ability [5-6].

The effectiveness of one-on-one instruction with respect to learning through dialogue is highlighted by researchers and educators [7-8-9]. However, little if any, is known about how reflective dialogues may lead to learning improvement and predict students’ problem solving ability. This information can be extracted from large educational data sets using data mining techniques. Consequently, this study aims at mining USNA physics data set applying clustering approach to find patterns in the data and identify how reflective dialogues predict students’ problem solving ability. The results of analysis help educators plan more effective tutorial dialogues in intelligent tutoring systems (ITSs). In sum, contributions of this paper are as follows:

a. By adopting clustering approach to mine educational data, this study contributes to better understanding of USNA Physics Data set and gives a better appreciation of reflective dialogues that might improve learning and predict students’ problem solving skills.
b. Two level (hybrid) clustering approach as used in this study amalgamates benefits of both supervised and unsupervised partitioning methods. The researcher is not required to have prior knowledge of every cluster; even so, by considering multivariate spreads precise covariance matrixes and accurate mean vectors can be obtained.
c. The results help ITSs developers design more efficient tutorial dialogues that improve students’ problem solving skills in a real classroom situation.

2. Literature Review

2.1 Post-practice Reflective Dialogues and Student Problem Solving

Researchers in the field of ITSs have recognized the importance of human tutorial dialogue in predicting student learning achievement [10-11]. During these dialogues, tutor gives students guidance on the problem and individual student’s solution to that particular problem. Clarifying the problem by dividing it into small pieces and demystifying how students could come up with the solution to the problem in an effective way, give tutors another chance to meet the learning needs of students.

Literature shows that many researchers have integrated post-practice reflective activities into ITSs. Reifying solution processes of individual students which sometimes was followed by feedback from an automated coach was one of these reflective tools [12]. Research works on the efficiency of reification of students' solution trace have indicated that such kind of reflective activities enhance self-assessment and help students perform better on the subsequent tasks [13]. Yet, in the aforementioned studies, little if any, attention have been paid to the importance of post-practice reflective dialogue between student and human tutor that may facilitate the implementation of other reflective tools [9] .

Previous research works in the context of live tutorials point to the effectiveness of post-solution reflection as a useful instructional activity. For instance, the study conducted by [14] on reflective dialogues in avionics revealed various kinds of student-tutor exchanges. [6] expanded the findings of the previous study and claimed that explanations distributed between problem solving and post-practice reflection are more efficient than problem solving explanations per se. They also asserted that post-solution reflections during live tutorials and ITSs differ in several ways. For example, post-solution reflection in live tutorials mostly revolves around specific errors in students' solutions and not narrative traces of their solutions. Moreover, post-solution reflections in live tutorials are more dialogic than in ITSs. Thus, much effort is needed to automatize more dialogic post-reflection discussions in ITSs.

A research work [9] investigated whether or not reflective questions and feedback on students' solutions enhance their conceptual knowledge and problem solving skills in Andes tutoring system. There were three conditions; the first treatment condition received reflective questions coupled with canned feedback and the second treatment condition was exposed to the same reflective questions but instead of receiving canned feedback they could interact with a human tutor. The control group solved Andes problems without receiving feedback and reflection questions. The results indicated that reflection questions both with human feedback and canned text feedback enhance learning. They claimed that this study is the first experimental study on post-practice reflection that illustrates its instructional effectiveness and value and more studies need to be conducted. However, this study did not demonstrate if the same results would be revealed in real classroom situation.

To this end, a follow up study was conducted by [15] to see if post practice reflective dialogues improve students' conceptual knowledge and problem solving ability in real classroom situation. Two experiments were conducted. The first experiment examined whether post reflective dialogues enhance leaning and also if students learn more when they interact during reflective dialogues.

Three reflective conditions with different activity levels were compared. In the first reflective condition, the least interactive version of reflective activity, students were provided with expert-generated feedback. This treatment condition known as canned-text response condition (CTR) did not lead to interaction but only self-explanation among students. In another reflective condition, KCDs reflective condition, knowledge construction dialogues (KCDs) were applied to guide students towards the correct response using Socratic Method of questioning. Yet, this condition did not give students a chance to ask questions. Therefore, in the third reflective condition, mixed initiative condition (MIC), following teacher turns there were hyperlinks that were related to questions students may intend to ask.

This experiment was conducted in fall 2005. 123 students in physics I classrooms took part in the experiment. They were randomly assigned to each condition. First, students had to take the pre-test. Then, the treatment groups solved the problem and answered the reflective questions (there were 9 work energy problems and 22 reflective questions). After the intervention all three groups sat for the post-test. In order to measure retention, an hourly exam was conducted at the end of work energy lesson.

The results indicated that students' level of interaction in problem solving and reflective dialogues was low so that it was difficult to conclude that students in more reflective condition outperformed those in less reflective condition. Before comparing the groups, data from those students who had a very low level of participation were omitted. Also, those students in reflective conditions who did no dialogue were considered as control subjects and reflective conditions of KCDs and MIX were combined due to the fact that only a few students in MIX condition asked probe questions. So, there remained 38 reflective dialogues, 17 CTR and 48 control subjects. Considering low participation level during problem solving and reflective dialogues, the results showed no significant differences amongst various reflective conditions on the post-test. However, the results revealed that students who were involved in reflective dialogues learn better than those who were not engaged in any reflective activity. It also indicated that the positive effect of reflective dialogues from the experiment hold up in the real classroom situation. This is in spite of the fact that this experiment failed to test the hypothesis that more reflective conditions could be more helpful than less interactive forms. For instance, there were only slight differences between canned text condition and other conditions in terms of final exam scores.

Therefore, [15] conducted another experiment in fall 2006 to refine the results of the previous experiment that was students who were involved in reflective dialogues performed better than other students as shown by the post-test. The current paper aims to mine this data set and all particularities are provided in section 3.

2.2 Clustering Approach

Clustering is an essential data mining tool for analysing and exploring educational data. Clustering is a prominent method to recognize new learning patterns and has been used in much recent research [16]. Hence, we describe briefly about clustering in the following.

The goal of clustering is assigning objects to groups that contain similar objects. Cluster analysis is a set of statistical methods that is widely used in several fields. Clustering approaches are based on maximizing the degree of association regarding the target variable in a group and minimizing it for members that belong to different clusters. Therefore, cluster analysis techniques enable researchers to organize large data sets and utilize them for the subsequent steps [16].

First step of cluster analysis is computing proximity indices among all members concerning the variable of interest. Whenever proximity indices are recognized then a clustering algorithm can be used to group similar data. Several clustering methods have been introduced but they are generally categorized into two groups: Hierarchical and Non-hierarchical.

Hierarchal clustering [17] is an approach of cluster analysis which makes a hierarchy of clusters using agglomerative or divisive algorithms. Agglomerative algorithm considers each item as a cluster, and then gradually merges the clusters (bottom-up). In contrast, divisive algorithm starts with all objects as a single cluster and then splits the cluster to reach the clusters with one object (top-down). In general, hierarchical algorithms are weak in terms of quality because they cannot adjust the clusters after splitting a cluster in divisive method, or after merging in agglomerative method. As a result, usually hierarchical clustering algorithms are combined with another algorithm as a hybrid clustering approach to remedy this issue. Moreover, some extended works are done to fulfil the performance of hierarchical clustering such as Chameleon [18], CURE [19] and BIRCH [20] where the merge approach is enhanced or constructed clusters are refined.

In the current paper, a two level clustering approach is used. First, hierarchical approach is used as the researcher does not have any accurate premise about the number of clusters in data set. Then, a partitional method is adopted when the number of clusters becomes evident.

A partitioning clustering method, makes k groups from n unlabelled objects such that each group contains at least one object. One of the most used algorithms of partitioning clustering is *k*-Means [21] where each cluster has a prototype which is the mean value of its objects. The main idea behind *k*-Means clustering is the minimization of the total distance (typically Euclidian distance) between all objects in a cluster from their cluster center (prototype). Prototype in *k*-Means process is defined as mean vector of objects in a cluster. However, when it comes to time-series clustering, it is a challenging issue and is not trivial [22]. Another member of partitioning family is *k*-Medoids (PAM) algorithm [17], where the prototype of each cluster is one of the nearest objects to the centre of the cluster. Moreover, CLARA and CLARANS [23] are improved versions of *k*-Medoid algorithm for mining spatial databases.

In both *k*-Means and *k*-Medoids clustering algorithms, number of clusters, k, is not available or feasible to determine for many applications and it has to be pre-assigned. So, it is impractical in obtaining natural clustering results and is known as one of their drawbacks in static objects [24]. However, *k*-Means and *k*-Medoids are very fast as compared to hierarchical clustering [25-21] and it has made them very suitable for clustering and has been used in several research works.

2.3 Clustering of Educational Data

Educational data mining (EDM) is considered a new discipline that is based on data mining techniques and algorithms and aims at exploring educational data to find predictions and patterns in data that characterize learners' behaviour [26]. One of the most useful EDM techniques is clustering approach. This approach has been used by several researchers in the field of EDM. For example, in their paper entitled "using cluster analysis for data mining in educational technology research" [27] used both hierarchical (Ward's clustering) and non-hierarchical (*k*-Means clustering) to analyse click-stream server-log data in order to find out the characteristics of learners' behaviour while they are engaged in problem solving in an online environment.

In their study [28] tried to combine unsupervised and supervised classification to build user models for exploratory learning environments. They used *k*-Means clustering approach to analyse logged interface and eye-tracking data with the aim of discovering and capturing effective and ineffective students' behaviours while interacting with exploratory learning environments.

Clustering approaches such as *k*-Means, classification (rule-based algorithms), tree-based algorithms, and function-based algorithms were used to analyse educational data. For instance, [29] analysed graduate students information from 1993 to 2007 using Association (Lift metric), classification (Rule-based and Naïve Bayesian), clustering (k-means) and outlier detection rules (Distance-based Approach and Density-based Approach) to improve graduate students' performance, and overcome the problem of low grades obtained by the students. [30] partitioned KDD Cup 1999 data set to proposes a hybrid model for intrusion detection to overcome difficulties with class dominance, force assignment and class problem. [31] classified data from 114 university students during a first year course in computer science using Classification (Rule-based algorithms, Tree-based algorithms, function-based algorithms, Bayes-based algorithms) and analysed the data using Weka/Clustering (EM, Hierarchical Cluster, SIB, K-Means), as well as association rule mining algorithm.

In order to identify key features of student performance in educational video games and simulations, [32] applied Fuzzy cluster analysis using "fanny" algorithm in R & "agnes" algorithm to partition log files generated by an educational video game called as "Save Patch".

Agglomerative hierarchical clustering algorithm was utilized in a study carried out by [33] to analyse Students' activity in time series form and determine what different behaviour patterns are adopted by students in online discussion forums. Another study by [34] applied clustering with latent class analysis (LCA) to group the Instructional Architect (IA) teacher users according to their diverse online behaviours in IA relational data set to understand teacher users of a digital library service.

Some studies [35] used clustering algorithm (expectation maximization) to analyse big educational data. In their study [35] applied this clustering algorithm to analyse data from 106 college students to distinguish different classes of learners based on performance and learning behaviours. [36] applied the same algorithm to Moodle forum used by university students during a first-year course in computer engineering to determine if student participation in the course forum can be a good predictor of the final marks for the course and to examine whether the proposed classification via clustering approach can obtain similar accuracy to traditional classification algorithms.

[37] analysed USNA physics spring 2007, 2008, and 2009 data sets using Clustering (Linear Discriminant Analysis) and sequential learning activity data to propose a new approach for the extraction of information from sequential user activities, and the analysis and interpretation of such information, with the ultimate goal of deriving Adaptation-oriented knowledge from naturally occurring learning behaviour.

As mentioned in passing, this study aims at partitioning USNA physics data set applying a two level clustering approach to find patterns in the data and identify how reflective dialogues predict students' problem solving ability.

3. The Corpus (Data set)

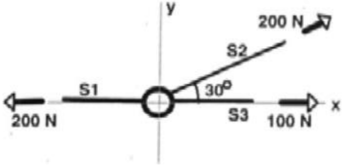
We have selected our corpus from Andes Physics data set taken from PSLC data shop; a previous study conducted by [15] on the effectiveness of reflective questions in students' problem solving when they answer reflective questions after solving physics problems with Andes tutoring system [38]. The experiment was conducted in first year physics classrooms at the US Naval academy. 67 students taking general physics I were recruited and randomly assigned to each group. Treatment group comprised of 33 subjects and there were 34 control subjects.

Students first took a pre-test consisting of 6 quantitative and 24 qualitative physics problems. It is worth mentioning that knowledge components (KCs) required to answer each problem were specified by physics experts. After the pre-test, participating students received training on how to use Andes. The intervention was carried out over the period of two weeks and covered 5 units of statics, translational dynamics, work energy, power, and linear momentum.

Upon completion of the problem, students in reflective group needed to answer a reflective question related to the problem they have solved. They typed the answer and started a teletyped dialogue with their tutor on the answer given. The dialogue between student and human tutor continued until the tutor ensured the student's understanding of the correct answer to the reflective question posed. Almost 3 to 8 reflective questions were asked per problem. Students could not proceed to the next problem until they accomplished all the reflective questions related to the problem. There were 26 problems all together for reflective group and control group solved 5 more problems to balance time spent on task. Reflective intervention comprised of 21 post practice reflective dialogue in all as well as 5 capstone dialogues (one dialogue at the end of each aforementioned unit). Figure 1, shows an example of reflective dialogues used in this experiment.

After the intervention, students took a post-test that was similar both in form and content to the pre-test. Before comparing across the groups, one of the subjects with post-test duration of less than 2 minutes was omitted from control group and 2 subjects who did not participate in dialogues were reclassified and considered as control subjects. Thus, there remained treatment and control group of 31 and 35 subjects, respectively.

Problem: In the figure below, each of the three strings exerts a tension force on the ring as marked. Use the labels S1, S2, and S3 to refer to the three strings. Find the components of the net force acting on the ring.



Reflection question: What if I now told you that this ring has an acceleration. If you knew the mass of the ring (3 kg), how would you solve for the acceleration?

Student: 73.2 _ 3_a; 100 _ Fw _ 3_a. Is this right; how would the acceleration be the same for both?

Tutor: You have to keep the a_x and a_y distinguished. They are two completely independent numbers that (together with a_z) specify your acceleration vector. You don't try to boil them down to one number. It's as if I told you, "To get to my house, you go 3 blocks north and 5 blocks east," and you said, "Ah, so you just go 8 blocks"—the two numbers together are the vector; they don't "boil down" to one number. OK?

Student: But can't it only have one acceleration?

Tutor: It does have only one acceleration, but that acceleration is a vector and it takes 3 numbers to write it down. You need to review vectors in some detail; a_x, a_y, and a_z together specify the acceleration vector.

Figure 1. Example of a Reflective Dialogue Between a Human Tutor and Student (Adopted from Andes Physics Tutor System Available at http://www.andestutor.org)

The results showed that reflective group performed better than control group on the post-test. Students' engagement both in reflective dialogues and problem solving considerably improved as compared to the previous experiment; yet, it was still far from perfection. The data showed that the positive results of reflective dialogues after solving Andes problems maintain in real classroom setting; however, it did not reveal a significant effect. It was presumed that capstone dialogues that reflective subjects were supposed to complete at the end of each unit before doing the Andes problems, would significantly enhance students' problem solving skills but the results did not show that significant impact of post-practice dialogues on students' problem solving ability. This might be because only a few capstone dialogues (5 dialogues) were given to reflective group to see any effect.

4. Findings

There were a total number of 345,536 transactions in all dialogues in reflective group. However, this data was still noisy because within the data set there were certain attributes whose values were inconclusive or useless. For example, 'Student response type' attribute had a value called "Choose" that was inconclusive so it was removed. There were 832 such records. Then, for the attribute of Duration there were 830 records with a value ‘dot’ which was again inconclusive. For Level (group) attribute there were 207 of 345446 records with value ‘* (asterisk)’. To remove these outliers, the data set was saved to comma separated values in IBM SPSS Modeler and then these outliers were manually removed.

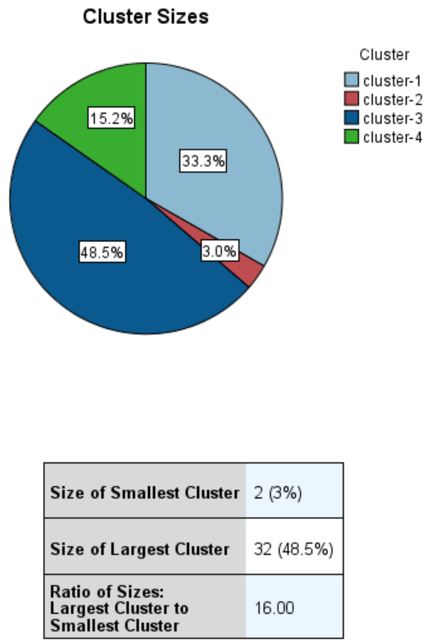
With the outliers removed, the dataset was reduced to 343588 instances and 14 attributes. We partitioned this data set with the following features (Table. 1). As mentioned in passing, in this study a two level clustering approach was applied. Since the researchers did not have any accurate premise about the possible number of clusters, hierarchical approach was used first. A non-hierarchical (k-Means) method was used afterwards.

TABLE 1. EXTRACTED FEATURES FROM DATASET

Feature	Meaning
Count of Session Id	shows how many times a student has started a new session
Sum of Duration (sec)	is the total time spent on problem solving activity
Avg of Duration (sec)	is the mean time spent on problem solving activity
Avg Count of attempts	shows how many time students make effort to find the correct answer to the problem
Avg Count of DELETION	shows the average number of times students delete their entries
Avg Count of CALC_REQUEST	indicates the average number of times students ask for calculator from their tutor during problem solving activity
Avg Count of HINT_REQUEST	shows the average number of times students seek help from their tutor during problem solving
Count of Problems	is indicator of the number of problems solved by each student
Avg Count of steps	shows the total number of steps taken by a student to solve a problem
Avg of correct	indicates average number of correct student entries during problem solving
Avg of Incorrect	indicates average number of incorrect student entries during problem solving
Condition Type	shows if the students belong to reflective condition or control group

Adopting two level (hybrid) clustering approach, we partitioned our feature dataset into four clusters. The pie chart (figure. 2) contains each cluster and the percentage size of every cluster is shown on each slice. As shown in figure 2, every cluster is of different size.

The percentage size of cluster 1 is 33.3% and the sizes of clusters 2, 3, and 4 are 3.0%, 48.5%, and 15.2%, respectively. As can be seen, cluster 2 with the percentage size of 3.0% is the smallest cluster and the largest one is cluster 3 with the percentage size of 48.5%. Also, the ratio of sizes (largest cluster to smallest cluster) is 16.00.



Size of Smallest Cluster	2 (3%)
Size of Largest Cluster	32 (48.5%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	16.00

Figure 2. A summary of cluster structure

Figure 3, indicates the predictor importance of our feature dataset. For example, “Condition Type” and “Average of Duration” are considered as two of the most important predictive features and “Count of Problems” as the least important one. This figure is important because it reveals which predictor matters most and also relates to the significance of each predictor in making estimations and predictions. All the features are explained in the following.

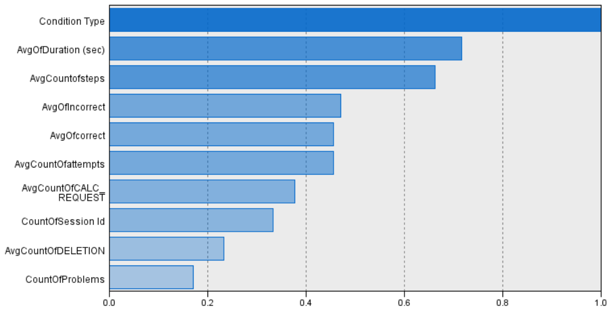


Figure 3. Predictor importance of feature in construction of cluster structure

Cluster 1 consisted of 22 students all of whom belonged to reflective condition (experimental group). Clusters 2 and 3 comprised of 2 and 32 subjects respectively and all of them belonged to control group. Lastly, the forth cluster consisted of 10 students from experimental group.

As for the next predictor, Average of Duration for clusters 1 and 4 was 16.64 and 15.07 seconds and for clusters 2 and 3 was 54.18 and 15.83 seconds, correspondingly. Cluster 2 took the longest to solve the problems. It can be said that cluster 2 subjects (DCBA5 & DD901) did not have a clear understanding of the problem in hand; therefore, they spent too much time to find the solution. Also, in cluster 3 there were some students who took very long to accomplish the problem given. To illustrate, DE141 with average duration of

20.73, DDBB9 with average duration of 24.03, and also DCBBD with average duration of 23.83 had the highest average of duration as compared to other subjects. All in all, control group spent more time on the problems than reflective group. The average of duration for control group was 16.14 which was considerably higher than that of reflective condition (14.91).

Average count of steps was different in each cluster. It was 15.36 and 16.70 respectively for cluster 1 and 4, 16.19 for cluster 3 and 5.50 for the smallest cluster (cluster 2). Average of steps for control group was 15.55 which is less than the Average of steps taken by reflective group (15.78). It is quite interesting to know that control subjects had pretty high average of duration during the problem solving activity but they took almost fewer steps to find the correct answers to the problems as compared to reflective subjects. Some of the control subjects had a very high average of duration but a very low average count of steps. For instance, DD901 who belongs to cluster 2, had a long average of duration of 82.25 second, yet, this student took only 4 steps to solve the problem.

The average of incorrect answers for cluster 1 was 22.55 and 31.10 for cluster 4. Moreover, 25.47 and 42.50 were the average of incorrect answers for clusters 3 and 2. It is clear that clusters 2 and 4 had the highest number of incorrect answers. For instance, average of incorrect answers for DD901 in cluster 2 was 50 which is relatively high. In cluster 4, DD391 had an average of 40 incorrect answers that was high as compared to other subjects in reflective condition. In general, the average of incorrect answers for control group was 26.47 which is significantly higher than the average of incorrect answers for reflective condition that was 15.5. This points to the fact that some control subjects with high average of duration, low average count of steps and high average number of incorrect answers did not have a good appreciation of the problems and just spent too much time on the questions without taking enough steps to find the correct answers.

The average of correct answers was pretty high for each cluster. It was 75.73 and 67.10 for clusters 1 and 4 respectively and 72.78 and 56.50 for clusters 3 and 2. However, as the results indicated, the average of correct answer for reflective condition (73.03) was higher than control subjects (71.82). The data showed that a few students have high average of both correct and incorrect answers (e.g., DD391 in cluster 4 with the average of 59 for correct answers and average of incorrect answers of 40 as well as DD901 in cluster 2 with the average of correct and incorrect answers of 50). It is implied that subjects with such kind of particularities were gaming the Andes tutoring system rather than putting effort to find the correct solution to the problems.

Another important predictor of students' problem solving ability was average count of attempts which was 21.95 for cluster 1 and 26.80 for cluster 4. 23.94 and 8.50 were respectively the average count of attempts for third and second clusters. The data showed that cluster 4 had the highest average of attempts amongst others. So, it can be inferred that average count of attempts in reflective condition was more than control group. The percentage of average count of attempts in reflective group was 23.46 and was much more higher that of control subjects with average count of attempts of 23.02. DD901 had the least average of attempts (Avg of attempts of 4) in control group and this is in spite of the fact that this subject spent a long time on the problems. Cluster 3 had a high average of attempts in control group; yet, it was still less than that of reflective condition.

The data also showed that students in cluster 2 did never request for calculator during problem solving. However, the average count of calculator-request among students in clusters 1, 3, and 4 was 1.81, 1.89, and 0.80, in the order mentioned. Cluster 3 subjects requested calculator from their tutor more than other students in other clusters. It was also found out that one of the students in cluster 3 (DDC3D) did not ask for calculator during problem solving. The average of calc-request for experimental group was 1.59 which was less than reflective subjects' request for calculator (1.89).

Moreover, DDC3D had no deletion during problem solving. The average count of deletion for cluster 1 and 4 was 1.09 and 2.3 correspondingly and 2.00 and 1.65 for clusters 2 and 3. The results revealed that the average of deletion among control subjects (1.67) was higher than that of reflective group (1.46).

As for the predictor of the count of session Id, cluster 4 subjects had the largest number of starting a new session (46.10). Count of session Id for clusters 1, 2, and 3 was 32.18, 12.00, and 41.16. One of cluster 4 students, DCC4D, and two of cluster 3 students, E04A3 and DD2D1, were found to have the highest count of session Id of 59. In addition, DD901 had started a new session 8 times only which was considered the lowest count of session Id. Generally, control subjects had a higher count of session Id of 39.44 which was relatively higher than that of reflective subjects (36.53).

Clusters number 4 and 3 with 140.90 and 137.72 counts of problem had the highest count of problems, respectively. Count of problems for cluster 1 was 121.14 and for cluster 2 was 66.50. DDE05 and DE057 from reflective condition with count of problems of 179 and 180 respectively were among the highest counts of problems. On the contrary, DD901 from control group had the lowest count of problems. Yet, the average count of problems in control group (133.52) was significantly higher than reflective group's average count of problems (127.31).

Cluster 4 with 698.64 had the highest sum of duration. Sum of duration for clusters 1, 2, and 3 was 531.06, 541.18, and 627.05, correspondingly. And finally, it was found out that cluster 3 had the largest average count of hint request (13.64). The average number of hint request for cluster 1, 2, and 4 was 9.56, 1.00, and 0.50 respectively. The data showed that control group requested more hints from their tutor than reflective subjects.

As mentioned earlier, reflective group performed better on the post-test. The result maintained in real classroom situation; however, it did not show a significant effect of reflective dialogues on students' problem solving. The results showed that experimental group with the average of 291.81 outperformed control group with the average of 286.41 on the first session of hourly exam. In the same way, on the next session of hourly exam, the average score for reflective subjects was 292.34 which is slightly higher than that of control group with the average score of 291.74.

It is also shown that both groups performed better on the second session of the hourly exam. It was revealed that some control subjects (e.g. DCBA5, DD901 in cluster 2 & DC71F, DC9C5, DC9E9, DE057, DE1CB in cluster 3) obtained high scores on the hourly exams despite the fact that they received no reflective question. Maybe this is due to the fact that these students were gaming the ANDES tutorial system. These students spent less time on the questions and asked for more hints from their tutor during the intervention. Table 2 provides a summary of our feature partitioning.

TABLE 2. DETAILS OF CLUSTER ANALYSIS

Descriptions	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Condition type	Experiment	Control	Control	Experiment
Avg of Duration	16.64	54.18	15.83	15.07
Avg Count of Steps	15.36	5.5	16.19	16.7
Avg of Incorrect	22.55	42.5	25.47	31.1
Avg of Correct	75.73	56.5	72.78	67.1
Avg Count of attempts	21.95	8.5	23.94	26.8
Avg Count of Calc-request	1.81	0	1.89	0.8
Count of session Id	32.18	12	41.16	46.1
Avg count of deletion	1.09	2	1.65	2.3
Count of problems	121.14	66.5	137.72	140.9
Sum of duration	531.06	541.18	627.05	698.64
Avg count of hint request	9.56	1	13.64	0.5
Avg of DT Exam scores 1	293.68	334	307	224
Avg of DT Exam scores 2	295.51	201	254	202

5. Concluding Remarks

From the above analysis we can infer that Andes was more of a procedural software system than an intelligent tutor system. In most of the logs the students have written explicit comments during their interaction with Andes. Andes would provide the students with a feedback that was pre-coded into it in case if the students' response was incorrect. Yet, there were no further explanations to the feedback.

The data from this study showed that the average of duration was longer in control group. Control subjects took lower average count of steps during problem solving activity. Also, higher average of correct answers and average count of attempts was found in reflective condition. Yet, control group had higher level of incorrect answers as compared to reflective group.

During the intervention, reflective subjects had the higher level of average count of attempts and lesser average count of deletion. Reflective group asked for less hints from tutor, had lesser count of problem and requested for calculator less than control subjects. The data also indicated that reflective subjects did not usually ask for too many hints, did not spent too much time on the problem but at the same time they had high average of correct responses.

In sum, in order to improve students' performance in ITSs, more reflective question needs to be asked of students. More importantly, they should not have the chance to find the correct answer by making conjecture.

References

[1] L. Vygotsky, *Mind in society: The development of higher psychological process*. Cambridge: Harvard University Press, 1978.

[2] S. Abbasnasab Sardareh and M. R. M. Saad, "Defining Assessment for Learning: A proposed definition from a sociocultural perspective," *Life Sci J*, vol. 10, no.2, pp. 2493- 2497, 2013.

[3] R. Berry, *Assessment for Learning*. Hong Kong: Kong University Press, 2008.

[4] S. Abbasnasab Sardareh, M. R. M. Saad, A. J. Othman, & R. Che Me, "ESL Teachers' Questioning Technique in an Assessment for Learning Context: Promising or problematic?" *International Education Studies*, vol. 7, no.9, pp. 161-174, 2014.

[5] S. Katz and D. Allbritton, "Improving learning from practice problems through reflection," in *the annual meeting of the American Educational Research Association*, 2002.

[6] S. Katz, G. O'Donnell, and H. Kay, "An approach to analyzing the role and structure of reflective dialogue," *Int. J. Artif. Intell. Educ.*, vol. 11, no. 3, pp. 320–343, 2000.

[7] K. E. Boyer, R. Phillips, A. Ingram, E. Y. Ha, M. Wallis, M. Vouk, and J. Lester, "Investigating the Relationship Between Dialogue Structure and Tutoring Effectiveness: A Hidden Markov Modeling Approach," *Int. J. Artif. Intell. Educ.*, vol. 21, no. 1, pp. 65–81, 2011.

[8] S. Katz, D. Allbritton, and J. Connelly, "Going beyond the problem given: How human tutors use post-solution discussions to support transfer," *Int. J. Artif. Intell. Educ.*, vol. 13, no. 1, pp. 79–116, 2003.

[9] M. Chi, M. Roy, and R. Hausmann, "Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning," *Cogn. Sci.*, no. 32, pp. 301–341, 2008.

[10] K. Forbes-Riley and D. Litman, "Investigating Human Tutor Responses to Student Uncertainty for Adaptive System Development," in *Affective Computing and Intelligent Interaction Conference (ACII)*, 2007.

[11] A. Ward, J. Connelly, S. Katz, D. Litman, and C. Wilson, "Cohesion, semantics, and learning in reflective dialogue," in *the Workshop on Natural Language Processing in Support of Learning: Metrics, Feedback, & Connectivity, held with the 14th International Conference on Artificial Intelligence in Education (AIED)*, 2009.

[12] S. Katz, A. Lesgold, E. Hughes, D. Peters, G. Eggan, M. Gordin, and L. Greenberg, "Sherlock 2: An intelligent tutoring system built upon the LRDC Tutor Framework," in *Facilitating the development and use of interactive learning environments*, C. P. Bloom and R. B. Loftin, Eds. Mahwah: Erlbaum, 1998, pp. 227–258.

[13] E. Wenger, *Artificial intelligence and tutoring systems: Computational and cognitive approaches to the communication of knowledge*. Los Altos: Morgan Kaufmann, 1987.

[14] C. P. Rose, "The Role of Natural Language Interaction in Electronics Troubleshooting," in *the Eighth Annual International Energy Week Conference and Exhibition*, 1997.

[15] S. Katz, J. Connelly, and C. Wilson, "Out of the lab and into the classroom: An evaluation of reflective dialogue in Andes," in *Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work*, R. Luckin, K. R. Koedinger, and J. Greer, Eds. Amsterdam: IOS Press, 2007, pp. 425–432.

[16] C. Li and J. Yoo, "Modeling student online learning using clustering," in *Proceedings of the 44th annual southeast regional conference on - ACM-SE 44*, 2006, p. 186.

[17] L. Kaufman, P. J. Rousseeuw, and E. Corporation, *Finding groups in data: an introduction to cluster analysis*, vol. 39. Wiley Online Library, 1990.

[18] G. Karypis, E. H. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer (Long. Beach. Calif).*, vol. 32, no. 8, pp. 68–75, 1999.

[19] S. Guha, R. Rastogi, and K. Shim, “CURE: an efficient clustering algorithm for large databases,” in *ACM SIGMOD Record*, 1998, vol. 27, no. 2, pp. 73–84.

[20] T. Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: an efficient data clustering method for very large databases,” in *ACM SIGMOD Record*, 1996, vol. 25, no. 2, pp. 103–114.

[21] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the fifth Berkeley symposium Mathematical Statist. Probability*, 1967, vol. 1, pp. 281–297.

[22] V. Niennattrakul and C. Ratanamahatana, “Inaccuracies of shape averaging method using dynamic time warping for time series data,” *Comput. Sci. 2007*, pp. 513–520, 2007.

[23] R. T. Ng and J. Han, “Efficient and effective clustering methods for spatial data mining,” in *Proceedings of the International Conference on Very Large Data Bases*, 1994, pp. 144–144.

[24] X. Wang, K. Smith, and R. Hyndman, “Characteristic-Based Clustering for Time Series Data,” *Data Min. Knowl. Discov.*, vol. 13, no. 3, pp. 335–364, May 2006.

[25] P. S. Bradley, U. Fayyad, and C. Reina, “Scaling Clustering Algorithms to Large Databases,” *Knowl. Discov. Data Min.*, pp. 9–15, 1998.

[26] S. Abbasnasab Sardareh, M. R. Mohd Saad, A. J. Othman, and R. Che Me, "Enhancing Education Quality Using Eduational Data," *Scholar Journal of Arts, Humanities, and Social Sciences.*, vol. 2, no. 3B, pp. 440-444, 2014.

[27] P. D. Antonenko, S. Toy, and D. S. Niederhauser, “Using cluster analysis for data mining in educational technology research,” *Educ. Technol. Res. Dev.*, vol. 60, no. 3, pp. 383–398, 2012.

[28] S. Amershi and C. Conati, “Combining unsupervised and supervised classification to build user models for exploratory learning environments,” *J. Educ. Data Min.*, vol. 1, no. 1, pp. 18–71, 2009.

[29] M. M. A. Tair and A. M. El-Halees, “Mining Educational Data to Improve Students’ Performance: A Case Study,” *Int. J. Inf.*, vol. 2, no. 2, pp. 140–146, 2012.

[30] K. K. Bharti, S. Shukla, and S. Jain, “Intrusion detection using clustering,” in *PROCEEDING OF ACCTA International Conference*, 2010, pp. 158–165.

[31] C. Romero, M. I. López, J. M. Luna, and S. Ventura, “Predicting students’ final performance from participation in on-line discussion forums,” *Comput. Educ.*, vol. 68, pp. 458–472, 2014.

[32] D. E. I. R. D. R. E. Kerr and G. K. W. K. Chung, “Identifying key features of student performance in educational video games and simulations through cluster analysis,” *J. Educ. Data Min.*, vol. 4, no. 1, pp. 144–182, 2012.

[33] G. Cobo, D. Garcia, E. Santamaria, J. A. Moran, J. Melenchon, and C. Monzo, “Modeling Students’ Activity in Online Discussion Forums: A Strategy based on Time Series and Agglomerative Hierarchical Clustering,” in *the 2nd International Conference on Learning Analytics and Knowledge*, 2012, pp. 248–251.

[34] B. Xu and M. Recker, “Understanding Teacher Users of a Digital Library Service: A Clustering Approach,” *J. Educ. Data Min.*, vol. 3, no. 1, pp. 1–28, 2011.

[35] F. Bouchet, J. M. Harley, G. J. Trevors, and R. Azevedo, “Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning,” *J. Educ. Data Min.*, vol. 5, no. 1, pp. 104–146, 2012.

[36] M. I. López, J. M. Luna, C. Romero, and S. Ventura, “Classification via clustering for predicting final marks based on student participation in forums.,” in *Proceedings of the 5th International Conference on Educational Data Mining*, 2012.

[37] M. Mirjam and A. Paramythis, “Activity sequence modelling and dynamic clustering for personalized e-learning,” *User Model. User-adapt. Interact.*, vol. 21, no. 1–2, pp. 51–97, 2011.

[38] K. VanLehn, C. Lynch, K. Schulze, J. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill, “The Andes physics tutoring system: Lessons learned,” *Int. J. Artif. Intell. Educ.*, vol. 15, no. 3, pp. 147–204, 2005.