The Effect of Question Ordering Using Bloom's Taxonomy in an e-Learning Environment

Dennis Castleberry
Center for Computation and
Technology Louisiana State University
Baton Rouge, United States of America

Steven R. Brandt
Center for Computation and
Technology Louisiana State University
Baton Rouge, United States of America

Abstract—This paper explores the ordering of questions which use Bloom’s Taxonomy in an intelligent tutoring system. In particular, it addresses three questions: (1) When questions are asked in forward order, is there a performance difference between paper-based and electronic methods? (2) Is there a difference between asking questions in forward and reverse order? (3) Finally, if the student answers a question incorrectly at a level, does the asking of additional questions at a lower level increase mean performance for the remainder of the test? To answer this, three experiments were conducted using an e-Learning system with students in introductory undergraduate computer science courses. Results indicate statistically significant performance differences between the paper and electronic media conditions as well as between the forward and reverse question order conditions. In addition, a marginally statistically significant difference due to asking “intervention questions” was found.

Keywords—Bloom’s Taxonomy, Intelligent Tutoring Systems, e-Learning, Program Comprehension

I. INTRODUCTION

Bloom's cognitive taxonomy organizes questions into levels depending on the cognitive functions required of the answerer. The levels are: knowledge, comprehension, application, analysis, evaluation, and synthesis. A brief overview is given below, with definitions and examples of questions covering the concept of one-variable equations (as they are first encountered in middle-school algebra):

- **Knowledge.** Recalling factual information. *What is an equation?*
- **Comprehension.** Assigning meaning to information; rephrasing in one’s own words. *Is 2+3 an equation? (If not, then explain why.)*
- **Application.** Applying a rule to a specific instance. *What is the value of y in the equation y = 2 + x if x = 4?*
- **Analysis.** Breaking information into parts and exploring relationships. *How are equations and expressions related?*
- **Evaluation.** Judging the use of knowledge or the validity of an argument. *What can equations be used for?*
- **Synthesis.** Utilizing knowledge to create a new solution to satisfy a goal. *Suppose a bag of apples costs $2 per apple in the bag plus a flat-rate charge of $1. Write an equation relating the number of apples in the bag (x) to the total cost of the bag of apples (y).*

Each category depends on the cognitive functions used in the previous category. That is, comprehension requires knowledge, application requires comprehension, and so forth. Furthermore the mastery of one of the levels is with respect to a given concept. A student may be able to synthesize solutions to problems dealing with expressions, but may not possess knowledge of equations, and thus could not solve problems involving equations.

The utility of Bloom's taxonomy lies in its ability to pinpoint the underlying cause of the student's problem-solving impasses [9]. Suppose a test of mastery of one-variable equations with the comprehension question “Is 2+3 an expression?” is given, and the student reaches an impasse. If the question “What is an expression?” is asked and the student does not know, the impasse can be attributed to a lack of knowledge about the definition of term "expression". If the student does know the definition yet nevertheless reaches an impasse, one might instead attribute it to a comprehension difficulty as such; which might be remedied by explaining the definition and giving examples of expressions, then continuing to test at the comprehension level.

Educators may have an intuitive notion of how to do this, but Bloom's taxonomy gives the ability to examine the impact of questions scientifically. By identifying the tested concept and the Bloom level of exercises, one can then form hypotheses about student responses to questions. One may wonder whether asking knowledge questions at the beginning of a test does in fact make a difference on responses to
comprehension questions later in the test. More generally a teacher may wonder: "What questions should I ask and in what order?"

Here is where we introduce the notion of an intelligent tutoring system. An intelligent tutoring system is a computer-based system which interacts with a student with the end-goal of teaching material, but adapts its interaction in some way to the student's responses, much as a human tutor would. With such a system, the scheduling of questions, collection of responses, and analysis of data can be automated; and, should a scientifically interesting result turn up, the system may even report it in a format which is ready for publication.

A. Previous Work

Bloom's taxonomy has been proven to be useful at the undergraduate level, and particularly in the field of software engineering [1][8].

EEG studies on students who solved computer problems whose Bloom levels were identified have shown that the levels may be distinctly clustered by EEG signals [4], implying differences in brain functions used by Bloom level. When students are told the Bloom levels of the problems they answer, they show a marked preference for higher-level problems [5][6].

While it can be used to pinpoint more intellectually satisfying problems, i.e. those at the higher Bloom levels, it has also seen success in program comprehension [2], where the asking of comprehension questions fosters code reading [7]. In addition, it has been useful for pinpointing the difficulties of novice programmers in a guided learning approach [9].

We have designed an intelligent tutoring system to explore the effect of question ordering as it pertains to Bloom's taxonomy. The system is based on previous research on executable paper systems (for generating publication manuscripts) which use R as an underlying framework [3]. Our intelligent tutoring system question bank is available for browsing on-line1. It currently supports short answer and multiple choice formats.

B. Research Questions

In this paper, we address the following questions:

- Q1: Does the immediate feedback on a schedule of questions ordered by Bloom levels make a difference in the student's overall performance?
- Q2: If questions are sorted by Bloom level, does the order matter? In particular is there a difference between forward and reverse order?
- Q3: Suppose questions are sorted by Bloom level. If a student answers a question incorrectly at a level, then does the asking of additional questions at a lower level improve performance in the remainder of the assessment?

II. METHODOLOGY

Our intelligent tutoring system has a bank of questions which are tagged by Bloom level, concept, and question difficulty. We can use our system to easily create groups of questions (which we call assessments) and issue them either on paper or electronically. Our intelligent tutoring system supports short-answer and multiple-choice formats, and can provide immediate feedback to the student on the correctness of their response.

A. Experiment 1

Our first experiment tested to see if there is a performance difference between computer-based assessment and paper-based assessment when questions are ordered by Bloom level. Our hypothesis was that students taking the computer-based assessment would fare better than those taking the paper-based assessment because of the immediate feedback offered by the computer-based assessment.

We designed a test of 10 questions (with 2 concepts, each concept having questions over 5 Bloom levels) to give to students2. The concepts tested were on recursion and binary trees, and were written to be language-independent. The questions were of multiple-choice and short-answer format. There were two knowledge, comprehension, application, analysis, and evaluation questions. No synthesis questions were asked because of the constrained formats allowed by the computer-based testing framework. An inter-rater reliability of 90% was determined by two independent raters, both computer science educators, who assessed the Bloom levels of the questions. After a point of disagreement about an evaluation-level problem, the test was adjusted to yield an 100% inter-rater reliability.

Evaluation-level problems are resistant to multiple-choice and short-answer formats because of the nature of the category. Our approach was to use multiple-choice questions of the "choose the best answer" format, in which there are many proposed uses of a concept or language construct, but one stands out as the most sensible from the vantage point of expertise.

The test resembled a quiz that might be given in the normal course of teaching the class. At the end of the quiz, the question "how satisfied were you with the (paper/ computer)-based medium?" was asked to gauge satisfaction differences as well. This question was answered on a Likert scale of 1 (not at all) to 7 (very satisfied).

1 The ITS we use can be browsed at the dedicated website: https://steam.cct.lsu.edu/assessment/

2 All questions may be viewed at the following dedicated website: https://steam.cct.lsu.edu/assessment/
For this experiment, volunteers were recruited from a Java programming class for introductory computer science students. All students volunteered; candy was offered as an incentive for all the experiments. We split the classroom into two groups, matched based on their current grade in the course. We gave the control group the paper quiz, and the experimental group the computer-based quiz.

An answer was scored as totally correct if it coincided exactly with the solution, and otherwise scored as incorrect. Correct answers were encoded with 1, incorrect with 0. A composite score was derived by summing these scores per-student.

B. Experiment 2

The second experiment tested the effect of ordering the questions by Bloom-level. For this experiment we designed another test of 10 questions (2 concepts, each concept having questions over 5 Bloom levels). This test also covered recursion and binary trees. In the control condition, questions were given in forward Bloom-level order. In the experimental condition, they were given in reverse order.

For both Experiments 2 and 3, participants were recruited in the same manner; however for this experiment a C++ class which followed the same concept track was also added to the pool. Matched-pairs were assigned to each group based on their current grade in each course. The test was constructed and inter-rater reliability gauged in the same manner, and the test was also scored in the same manner.

C. Experiment 3

The third experiment tested the effect of intervening questions on the performance of later questions in the assessment. Our hypothesis was that overall performance would be improved if incorrect answers triggered the addition of new intervention questions from a lower Bloom level.

Experiment 3 participants were recruited from a different class. The test was this time language-dependent (MATLAB) and tested mastery of for-loops. In the control condition, the control group was given an assessment of 10 questions, with 2 questions per the first five Bloom levels. The experimental group was given an adaptive measure. If at any point a student answered a question incorrectly, then a question at the next lowest level was given. This applied to all levels except knowledge. So for example if a student answered an application-level question incorrectly, a comprehension-level question (related to the application-level question) was scheduled before another application-level question of the same type. The experimental group thus had a maximum of 4 additional questions asked for a total possible 14-question test.

III. ANALYSIS AND RESULTS

A. Experiment 1

The experimental condition \((N=27, M=6.21)\) did in fact show a higher mean score than the control condition \((N=27, M=5.23)\) in overall performance. Statistical significance was tested with a one-tailed two-sample matched-pairs Student's t-test on the composite score. The result indicated a statistically significant difference \((t=2.024, \ p=0.048)\). The effect size was large for this experiment (nearly one point). The experimental condition \((M=4.93)\) showed a higher mean score than the control condition \((M=4.38)\) in satisfaction as well; a similar t-test was done and was marginally statistically significant \((t=1.7753, \ p=0.082)\), however the effect size was small (.55 points).

B. Experiment 2

The experimental condition \((N=48, M=4.94)\) showed a higher mean score than the control condition \((N=48, M=4.31)\). Statistical significance was tested with a one-tailed parametric Student's t-test on the composite score. The result indicated a statistically significant difference \((t=2.13, \ p=0.036)\). However, the effect size was smaller than for the previous experiment (.63 points).

C. Experiment 3

To tell the immediate effect of the intervention questions, one-tailed parametric Student's t-test on the composite score of questions starting after the first intervention question was done. It was hypothesized that the experimental condition would perform better on the remainder of the test. The experimental group \((N=45, M=6.98)\) outperformed the control group \((N=45, M=6.23)\). The result indicated a marginally statistically significant difference \((t=1.71, \ p=0.092)\). The point difference was medium (.75).

IV. LIMITATIONS

A few validity concerns are to be pointed out. No random order was given in Experiment 2 because of a lack of available subjects; hence it cannot be inferred that forward order is no different from random order. The experimenters (paper authors) designed the tests. The number of items per test was small to allow for a conservative testing time. For Experiment 1, confounding variables (those other than immediate feedback) may have played a role in test-taking because the test-taking media were different.

Furthermore it is worth noting that while the mean differences in the case of all experiments were statistically significant, the effect size for Experiment 2 was small.
compared to Experiment 1 (which had an effect size of nearly one point).

IV. CONCLUSIONS

Results of the experiments support the hypotheses that question scheduling has an effect on performance. When questions are in Bloom forward-order, immediate feedback appears to have an effect on success of later questions. Scheduling questions in forward Bloom order in particular appears to have an advantage over the reverse. Finally, there is some preliminary evidence to suggest that asking intervention questions in response to incorrect answers may improve composite performance on a target subset of later questions.

These findings have interesting implications for e-Learning systems which use Bloom's taxonomy, in particular regarding question schedules. Future work will investigate more dynamic approaches to scheduling questions with the aim of improving aggregate performance on all questions, as well as improvements to the e-Learning system to handle the dynamic scheduling of questions.

REFERENCES